2P1-3

ユーザの特徴抽出による マルチエージェントモデルを用いた意見の自動評価

Automatic evaluation of opinions using multi-agent model and users' feature extraction

井田牧子 *1　藤田桂英 *1
Makiko Ida　Katsuhide Fujita

*1東京農工大学大学院　工学府　情報工学専攻
Department of Computer and Information Sciences, Graduate School of Engineering, Tokyo University of Agriculture and Technology

It is effective to evaluate users' opinions on online discussion bulletin board system using collective intelligence. For example, Mark et al. proposed the "Bag of Lemons" which is a type of crowd-based filtering evaluation method by general users not experts in open innovation systems. Considering crowd-based filtering evaluation method, we propose the multi-agent based approach for evaluating the opinions on electronic discussion bulletin board automatically. In our proposed method, the users' interests of topics and preferences are extracted from user's comments to generate the agent models. After that, our proposed method evaluates opinions automatically by social choice methods using the extracted multi-agent model. Experimental results show that our proposed multi-agent based filtering method is the most effective by filtering using Bag of Lemons based method, which is filtered by negative comments than other methods.

1. はじめに

近年，Web 技術の発展に伴い，SNS，電子掲示板など誰もが手軽に意見を述べることのできる多種多様なソーシャルメディアが登場している．これらのソーシャルメディアにおける投稿は時間や場所にとらわれず，膨大な量の情報を目にすることが容易となっている，一方，投稿数が膨大な場合，すべての内容に目を通し，内容を吟味した上で情報を取捨選択することが困難である．そこで，ユーザの意見や考えを自動抽出し，投稿意見を自動評価することは重要となる．

Mark らはオープンイノベーションシステムにおいて，専門家でなく一般ユーザからの評価を利用したアイディア選別手法，crowd-based filtering 手法の一種である "Bag of Lemons" という手法を提案している [1]．オープンイノベーションとは，組織や国を超え広く一般から意見を集めるという概念であり，近年多くの企業や民間団体などが，web 上でアイディアを募るオープンイノベーションシステムを活用している．オープンイノベーションシステムでは膨大なアイディアが集まるため，アイディアの選別が必須である．そこで，Mark らの提案手法では，複数人の一般ユーザにクライアントの評価基準と定められた数のポイントが与え，優れたアイディアとして選ばれなさそうなアイディアに，ポイントを分配する手法である．大規模な実験により，すべての意見に評価値を与える手法（Likert）と優れたものにポイントを分配する手法（Bag of Stars）に比べ，精度が向上されたことを示している．

本論文では Mark らの手法をもとに，議論掲示板から効用の高い投稿意見を自動で抽出することを目的とする．提案手法では，マルチエージェントモデルと単純な Bag of words により，エージェントモデルの生成と意見の自動評価を行う．まず，LDA と BM25 によりユーザのコメント群より各ユーザの興味情報を抽出する．次に，単語の評価極性により，各ユーザの選好情報を取得する．その後，作成したマルチエージェントモデルに対して，現実の群衆によるアイデアの選定に用いられる手法（Bag of Lemons, Bag of Stars, Likert）を適用し，投稿意見の自動評価を行う．

以下に，本論文の構成を示す．まず，本論文で実験対象とした議論掲示板の詳細と意見評価に関する既存研究を示す．次に，ユーザの特徴抽出を利用した意見評価手法を提案する．その後，評価実験と結果の議論を行い，本論文のまとめを示す．

連絡先: 井田牧子，東京農工大学工学府情報工学専攻，東京都小金井市中町 2－24－16，makiko.ida@katfuji.lab.tuat.ac.jp

2. 関連研究

2.1 大規模議論のためのプラットフォーム

大規模議論のためのプラットフォームとは，Web システム上で大規模な議論を行うことに特化したシステムである．本論文では，COLLAGREE[2, 3] と SYNCLON[4] という最終的に合意を得ることを目的とした，大規模議論のためのプラットフォームを対象に評価実験等を実施する．以下に各システムの詳細を述べる．

COLLAGREE

COLLAGREE とは，名古屋工業大学を中心として開発されている大規模意見集約システムである [2, 3]． COLLAGREE では，議論において中立な立場を保ち，意見集約を適切にリードする役割をもつファシリテータを中心に参加者が Web 上で議論を行う．議論の発散・整理・集約フェーズすべてに関する支援機能を提供し，大規模な人数での意見集約を効果的に支援することを目的としている．また，社会実験として，名古屋市時期総合計画に関する議論テーマについて，ファシリテータの専門家を交えた議論を行っている．

SYNCLON

SYNCLON とは株式会社ベクトカルチャー (Vectculture) が運営しているコミュニケーションサイト，議論掲示板である [4]．価値観の可視化・分類を通し，他人と価値観を共感しつつ議論を行うことで，前向きで建設的な議論ができる環境を構築していくことを目的としている．価値観は大まかに「社会，自由，美徳」の三つに分類されており，議論テーマは政治から文学まで様々である．

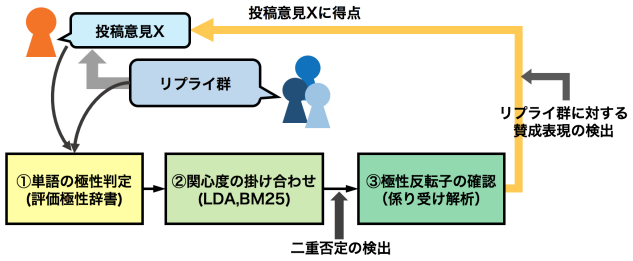

図 1: 投稿意見の評価手順

2.2 意見の評価と評価極性

自然言語文から評価情報を自動抽出する研究の中に，評価表現辞書の構築が存在する．評価表現辞書の構築は主に，語彙ネットワークを利用する手法，共起情報を利用した手法，周辺文脈の情報を利用した手法が用いられる．以下に，本論文で使用した日本語評価極性辞書の詳細を示す．

【日本語評価極性辞書 (名詞編，形容詞編)】

日本語評価極性辞書とは東北大学の乾研究室において開発されている．用言編は小林らが共起情報をもとに収集した評価表現約 5,000 件のリストを一部改編し，人手で評価極性情報を付与したデータである [5]．名詞編は東山らがポジティブな名詞と述語は共起しやすいという評価極性についての述語の選択選好性を利用し，約 8,500 の評価極性を持つ名詞表現に対して評価極性情報を付与したデータである [6]．用言編・名詞編ともに，ポジティブ，ネガティブ，ニュートラルで単語を評価している．

【日本語アプレイザル評価極性辞書】

日本語アプレイザル評価極性辞書 (JAppraisal 辞書) とは，評価表現を肯定的か否定的かだけではなく，評価基準（愛情、倫理に関する基準など）によって分類した電子化辞書である [7]．国立国語研究所コーパス開発センター (GSK) が 2011 年に公開，評価表現に該当する 8,544 語収録しており，総計 277 カテゴリに評価表現を分類している．

3. ユーザの特徴抽出に基づいた意見評価手法

前処理

まず，形態素解析や構文解析の精度を下げないために，文の一般化・平滑化を行う．具体的にはカタカナと英数字の表記の統一や，文中スペースの削除，また，URL と鉤括弧内の文は正規表現により削除を行う．

ユーザの特徴抽出によるエージェント効用モデル生成

各ユーザからエージェントの効用モデルを生成するために，まず議論内の単語をトピックに分類する．単語は名詞，形容詞，形容動詞，動詞を対象とし，ストップワードリスト [8] を参考に，トピック抽出に余分な単語は取り除いた．各投稿意見を一つの文書とみなして，トピックモデル LDA[9] を用いて単語をトピックに分類した．本論文では，予備実験の結果よりトピック数は 10 に設定した．名詞群をトピックに分類した後，各ユーザのトピックに対する関心度および効用を数値化する．各ユーザの発言群を一つの文書とみなし，名詞，形容詞，動詞に対して BM25[10] を用いて関心度を数値化した．また，ユーザの発言数が少ない場合は特徴量の抽出が正確にできないと判断し，発言数が 5 以下のユーザは対象外とした．

抽出した各エージェントの関心度と評価極性を用いて，それぞれの投稿意見の評価を行う．意見の評価の概要を図 1 に示す．まず，文中の単語の極性を確認し，ポジティブな単語は +1，ネガティブな単語は −1 とする．その後，その単語が属するトピックの関心度を掛け合わせる．ここで,「諦めるべきではない」といったように，その単語自身の評価極性が反転している場合が多く存在するため極性反転を抽出する．乾らは評価極性を反転させ，ポジティブな単語と組み合わさるとポジティブ値を示し，ネガティブな単語と組み合わさるとネガティブ値を示す単語を pn 演算子として定義している [11]．係り受け解析を用いて，その単語に係っている文節，もしくは係り先の文節に pn 演算子があり極性が反転する場合は，−1 をかけることにより極性を反転させる．

しかし，ユーザ群 U 内のあるユーザ u_k の各トピックに対する関心度の配列を I_{u_k}，係り受け解析により抽出した pn 演算子の値を ope，単語 w の評価極性の値を $p\_dic[w]$ としたとき，ある投稿 (意見)o_i に対するスコアは式 (1) で表される．そして，リプライ群のスコアを反映した評価値は式 (2) で表される．リプライ群のスコアを投稿 (意見)o_i に反映させる場合は，そのリプライ数で割る．

$$score(o_i, u_k) = \sum_{w \in all\ word\ of\ o_i} I_{u_k}[w's\ t\_num] * ope * po\_dic[w] \quad (1)$$

$$evaluation(o_i, u_k) = score(o_i, u_k) + \frac{1}{num\ of\ Rep(o_i)} \sum_{r \in Rep(o_i)} score(r, u_k) \quad (2)$$

以上の処理に加え，二重否定表現は極性反転のみで対応が難しいため，林らの論文 [12] をもとにテンプレートマッチングで極性を決定した．また,「そうですね」などの賛成語句が出現した場合はリプライ先の意見にポイントを付与するなどの処理を行った．

マルチエージェントモデルによる意見の自動評価

Mark らの論文をもとに下記の 5 通りで投稿意見の評価を行う．各エージェントの評価得点の合計点を投稿意見の評価値とする．

- **Likert_Tp_I (トピックモデル+関心度)**
  Likert 方式は，評価者が全ての意見を何段階かで評価し，その評価を足し合わせる方式である [1]．各エージェントの評価値を足し合わせ，その合計値により投稿意見をランキングを決定する．

- **Likert_I(関心度)**
  Likert 方式では単語をトピック分割して各ユーザのトピックごとに関心度を抽出していたが，単語ごとに関心度を抽出する．その後，Likert 方式と同様に意見を評価し，ランキングを決定する．

- **Likert_Tp(トピックモデル)**
  Likert 方式の比較として，各ユーザの関心度を抽出するフェーズを省略する．その意見に出現する単語の極性のみで意見を評価する．

- **Bag of Stars(BOS)**
  Bag of Stars は，全評価者が複数ポイントを良いと思う

表 1: COLLAGREE データ

議論 No.	議論テーマ	参加者数	コメント数
C_1	名古屋市時期総合計画 (災害)	20	332
C_2	名古屋市時期総合計画 (魅力)	34	392
C_3	名古屋市時期総合計画 (環境)	20	261
C_4	名古屋市時期総合計画 (人権)	20	238
C_5	AICHI 街づくり	40	350

意見に分配し，全評価者の合計ポイントで意見を評価するアイディアフィルタリング手法の一つである [1]．ポジティブな評価のみを抽出し足し合わせ，その合計値により投稿意見のランキングを決定する．

・Bag of Lemons(BOL)

Bag of Lemons は，Bag of Stars 方式に対して，ネガティブな評価のみをカウントし，意見の評価を行う．

4. 評価実験

4.1 実験設定

本研究の実験データとして，COLLAGREE[3] と SYNCLON[4] で実際に発言されたデータを用いる．表 1 に COLLAGREE データの詳細，表 2 に SYNCLON データの詳細を示す．

表 2: SYNCLON データ

議論 No.	議論テーマ	参加者数	コメント数
S_1	日本の国民食 (R 議論, 合意形成ゲーム)	3	118
S_2	ロボット/人工知能の未来についての議論	8	149
S_3	3D プリンターで拳銃製造について考える	7	75
S_4	国際社会は世界共通語を導入すべきである	6	93
S_5	夫婦別姓	6	108

今回，提案手法の評価を行うために，学生 5 人により実験データセットの全コメントの評価を行った．評価基準を下記の 3 項目のように定めた．

1. テーマに沿っているか
2. 参加者に興味や同意、共感を得られているか
3. その後の議論の発展に寄与しているか

評価方法は $[bad, average, good, excellent]$ の 4 段階と NA(判定不能) で評価した．ただし，評価が $average$ に極度に偏るのを防ぐため，各評価を最低 10% はつけるように促した．

4.2 実験結果と考察

提案手法の性能を評価するため，2 クラスの分類における分類器（分類アルゴリズム）の性能の評価指標として ROC 曲線と，ROC 曲線の下側の面積である AUC(Area Under the Curve)[13] を用いる．実験は，5 人の学生により作成された正解データを「$bad = 0pt, average = 1pt, good = 2pt, excellent = 3pt$」とし，平均評価値が 2pt 以上である意見を抽出することに対する指標を評価した．

表 3: AUC による各手法の比較 (COLLAGREE)

	C_1	C_2	C_3	C_4	C_5	Average
Likert_Tp_I	0.47	0.5	0.58	0.27	0.48	0.433
Likert_I	0.5	0.48	0.57	0.29	0.46	0.427
Likert_Tp	0.32	0.36	0.45	0.24	0.51	0.365
BOS	0.44	0.45	0.61	0.25	0.5	0.428
BOL	0.62	0.58	0.53	0.54	0.47	0.525

表 4: AUC による各手法の比較 (SYNCLON)

	S_1	S_2	S_3	S_4	S_5	Average
Likert_Tp_I	0.84	0.6	0.52	0.47	0.43	0.572
Likert_I	0.83	0.61	0.4	0.47	0.47	0.556
Likert_Tp	0.83	0.57	0.49	0.54	0.5	0.586
BOS	0.73	0.54	0.35	0.4	0.47	0.498
BOL	0.71	0.62	0.66	0.51	0.5	0.60

表 3，4 に本実験の AUC 値の一覧をまとめる．また，ランダム値 0.5 以上の場合を赤字で示す．

表 3, 4 の結果から，Bag of Lemons が最も性能が高い結果となった．しかし，議論データ S_1 と S_2 はほぼすべての手法で 0.5 以上であり，各手法とも議論データに性能が大きく依存している．Bag of Lemons が最も高精度となる結果は，Mark らのオープンイノベーションシステムにおける人手によるフィルタリング精度結果と同様の理由が考えられる [1]．Mark らは Bag of Stars より Bag of Lemons の方が高精度である理由として，人間は優れたものよりそうでないものを判断する方が容易であるためではないかという仮説を立てており，本実験の議論データにおいても，同様の仮説が成り立っている．また，Bag of Lemons の AUC の結果の分散が最も小さいことからも，ネガティブ評価のみを抽出する手法は，議論データの種類や話題に影響されにくいと考えられる．

次に，表 2 に各手法による意見評価において 0pt 評価となった意見の含有率，表 5 に各議論データにおける評価極性をもつ単語の出現回数を示す．

提案手法全体の課題として，0pt の意見が多いことが挙げられる．表 5 からわかるように，議論データ全体として評価極性の単語の出現率の低い．提案手法では，単語に評価極性がない場合，意見の評価に反映できず，精度は評価極性辞書の単語の出現率に依存してしまい，このことが多くの意見に対して 0pt

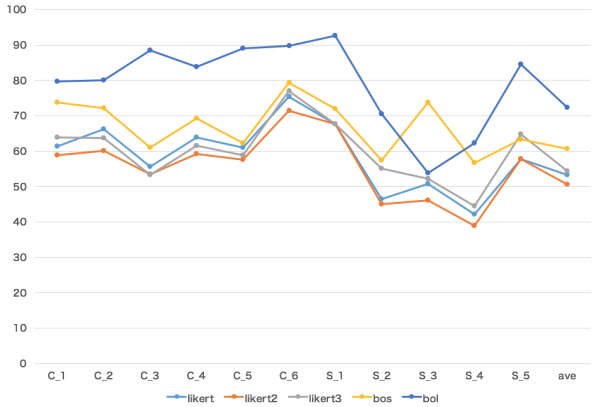

図 2: 0pt コメントの含有率

表 5: 評価極性の単語の出現率

議論 No.	ポジティブ	ネガティブ	pn 演算子 (マイナス)
C_1	129(3.5%)	30(0.8%)	30(0.8%)
C_2	191(3.6%)	34(0.6%)	22(0.4%)
C_3	388(4.9%)	46(0.6%)	51(0.6%)
C_4	126(3.2%)	15(0.4%)	13(0.3%)
C_5	206(4.0%)	24(0.5%)	35(0.7%)
C_6	136(3.2%)	87(2.1%)	23(0.5%)
S_1	77(7.9%)	0(0.0%)	8(0.8%)
S_2	129(3.5%)	30(0.8%)	30(0.8%)
S_3	101(5.2%)	8(0.4%)	17(0.9%)
S_4	80(3.0%)	19(0.7%)	27(1.0%)
S_5	48(2.8%)	14(0.8%)	6(0.4%)

の評価を引き起こしている．

表 3,4 における AUC 値による評価では，議論データごとに精度が偏っており，COLLAGREE と SYNCLON では SYNCLON のデータにおける精度の方が高い結果となった．特に，SYNCLON の S_1 における提案手法の精度は全般的に高い．この議論ではファシリテータが「発散タイム → 合意形成タイム → 合意タイム」という流れを事前に提示し，最終的に 1 つの合意案に全員が賛成しており，今回扱った他の議論データでは見られない大きな特徴である．一般的に，SYNCLON の S_1 のように，1 つの合意案に収束した議論では，何らかの案に対し「いいですね，賛成です」といったポジティブな極性を持つ単語が多く出現し，そのリプライ先の意見が評価されている場合が多い．以上の傾向から，Likert 手法や Bag of Stars の精度が高くなっている．COLLAGREE は SYNCLON に比べ，議論が収束せずにスレッドごとに分散する傾向がある．COLLAGREE のような議論では，各エージェントごとにユーザ情報によって評価する意見を限定する必要がある．COLLAGREE は SYNCLON と比較すると倍以上の議論ユーザがいる一方で，各ユーザのコメント数が少ない．本実験では，コメント数が 5 以下のユーザは対象としなかったが，平均コメント数などによって閾値を変化させるなどの対応も必要である．

提案手法の精度が低い原因の一つとしては，各コメントに対するリプライ数の少なさが挙げられる．Bag of Lemons を提案している Mark らがデータとして利用したオープンイノベーションシステムは，その意見に対する賛成意見と反対意見が少なくとも 3 つずつ以上掲載されており，それらを評価に反映することができる．一方，議論データではリプライがない場合，そのコメント自身の文書のみで評価しなくてはならない．しかし，そのコメント自身の極性情報が直接そのコメントの効用値に関連するとは限らないため，リプライ数が 0 のコメントが正しく評価されない可能性が大きい．以上から，ユーザ情報，例えばそのユーザのコメント数や影響力を意見の評価に反映させる，または，コメントのクラスタリングによってそのクラスタの情報を評価に反映させるなどの改善が必要である．

5. まとめ

本論文では，大量の投稿を含む議論掲示板から，効用の高い意見を自動で抽出することを目的とし，議論掲示板において，参加ユーザの意向を反映した意見の抽出を行う手法を提案した．また，議論掲示板 COLLAGREE と SYNCLON の議論データに対して，提案手法の評価実験を行った．実験の結果，評価の精度は高くはなかったが，ネガティブの評価値を抽出し，意見の評価を行うことが有効であることが確認された．

今後の課題としては指示語補完や，単語ではなく文極性の判定などが挙げられる．また，COLLAGREE などの議論掲示板で，各ユーザに対するお勧めコメントの提示や議論の構造化に応用することも考えられる．

謝辞

本研究は，JST，CREST の支援を受けたものである．

参考文献

[1] Mark Klein and Ana Cristina Bicharra Garcia: High-speed idea filtering with the bag of lemons. Decision Support Systems, Vol. 78, pp. 39-50, 2015.

[2] 伊美裕麻, 伊藤孝行, 伊藤孝紀, 秀島栄三: ファシリテータ支援機構に基づく大規模意見集約システム collagree の開発と評価名古屋市次期総合計画のネット上のタウンミーティングでの社会実験. 情報処理学会第 76 回全国大会, 2014.

[3] 伊美裕麻, 伊藤孝行, 伊藤孝紀, 秀島栄三ほか: オンラインファシリテーション支援機構に基づく大規模意見集約システム collagree-名古屋市次期総合計画のための市民議論に向けた社会実装-. 情報処理学会論文誌, Vol. 56, No.10, pp.1996-2010, 2015.

[4] 株式会社 Vectculture: 価値観の対話場シンクロン. http://synclon3.com/.

[5] 小林のぞみ, 乾 健太郎, 松本裕治, 立石健二, 福島 俊一: 意見抽出のための評価表現の収集. 自然言語処理, Vol.12, No.3, pp.203-222, 2005.

[6] 東山昌彦, 乾健太郎, 松本裕治: 述語の選択選好性に着目した名詞評価極性の獲得. 言語処理学会第 14 回年次大会論文集, pp.584-587, 2008.

[7] 佐野大樹: 『日本語アプレイザル評価表現辞書（態度表現編）』の構築. 言語処理学会第 17 年次大会, pp.115-118, 2011.

[8] オープンソース：日本語ストップワード, http://svn.sourceforge.jp/svnroot/slothlib/CSharp/Version1/SlothLib/NLP/Filter/StopWord/word/Japanese.txt, 2017.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan: Latent dirichlet allocation. Journal of Machine Learning Research, Vol.3, pp.993-1022, 2003.

[10] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze: An Introduction to Information Retrieval, Cambridge University Press, p. 233, 2009.

[11] 乾孝司, 乾健太郎, 松本裕治: 出来事の望ましさ判定を目的とした語彙知識獲得. 言語処理学会第 10 回年次大会発表論文集, pp.91-94, 2004.

[12] 林楽常: 二重否定表現の一考察: 形式と意味の相関性を中心に. 人間文化研究, Vol. 3, pp.27-39, 2005.

[13] Anthony K Akobeng: Understanding diagnostic tests 3: Receiver operating characteristic curves. Acta Paediatrica, Vol.96, pp.644-647, 2007.