1C2-R-5-1

議論能力に基づくWikipediaにおける編集者間議論ページの分析

Analysis of Discussion Page in Wikipedia based on User's Discussion Ability

朱 成敏[*1*2] 武田 英明[*2*1]
Sungmin JOO Hideaki TAKEDA

[*1]総合研究大学院大学 Graduate University for Advanced Studies
[*2]国立情報学研究所 National Institute of Informatics

Discussion page play a fundamental role in Wikipedia as the place for discussion and communication. In this study, we focus on communication patterns and debating abilities that accompany collaboration and propose the method to measure discussion quality in Wikipedia. We show the structure model and evaluation by experiments.

1. はじめに

近年 BBS などのツールを利用して様々な議論がインターネット上で活発に行われている．特に Wikipedia のように共同コンテンツの作成など協調作業が要求される場合, 議論の役割は大きい.

しかし, Wikipedia で行われている議論の中では記事の修正や編集方針について編集者達がお互い共感を得られる結果を出すため議論する, いわゆる建設的な議論も存在する一方, 結果が出せずに編集者間の相互誹謗により無意味な論争が長く続く非効率な場合も見られる．こういった場合, 建設的な議論への修正のために, その議論が結論形成において妥当性のある議論なのか判定することが必要となる.

そこで, 本稿では議論を行うに当たり要求される参加者の能力をスキルとして定義し, 参加者の発言からその能力を判断する．そして, 評価された参加者の議論能力が議論全体の妥当性の評価に与える影響と関係性を把握し妥当性の判定手法として提案, 実験・評価を行う．また, 判定基準をモデル化し, 参加者と議論の自動評価への可能性について述べる.

2. 関連研究

Wikipedia の参加者評価に関する研究は更新履歴や編集者の執筆成果を重視する研究 [1][2] が主に行われている．Wikipedia 記事の信頼度に関する研究は, 編集履歴を分析し, 削除なく残っている部分を作成した人の信頼度を上げ, また信頼度が高い編集者が参加した記事に高い信頼度を与える．すなわち執筆成果によって編集者を評価して, その編集者の信頼度から執筆した記事の信頼度を評価する手法が中心となった．これらの研究は「信頼度が高い編集者が参加した記事は信頼度が高い」という前提から記事を評価した.

インターネット上で行われる議論を分析した既存研究は特徴表現の抽出による重要度の判定や話題の流れから議論を可視化する研究 [3][4] が中心として行われてきた．構文分析によって重要単語を抽出して論点を定義した．論点から発言の繋がりを算出し, そのデータを中心として流れを可視化した．これらは説得システムとしての議論が持つ特徴よりはテキストの特徴抽出に注目した研究である．本研究では議論参加者の能力を定義し, その特徴から議論を評価する手法を提案する.

連絡先: 朱 成敏, 総合研究大学院大学複合科学研究科, 〒 101-8430 東京都千代田区一ツ橋 2-1-2, joo@nii.ac.jp

3. 議論の妥当性モデル

本章では先行研究と議論の妥当性を判定するための議論の特徴 (feature) に関して述べる.

3.1 議論の妥当性

論証における妥当性は導き出された結論の真偽ではなく結論形成までの形式によって判定される [5]. 従って議論の妥当性の判定も結論形成の過程に対する判定と考えられる．本研究では議論の進行や発言など参加者の姿勢から議論の妥当性の推測を目指す.

3.2 先行研究

先行研究 [9] として Wikipedia の議論データを対象として議論ページの評価を行った．議論の評価項目として論証能力とコミュニケーション能力に関する 4 つのスキル (自己主張, 他者受容, 関係調整, ディベート) から 6 つの評価項目を提案した．提案した評価項目をアンケート実験により議論データの評価を行った．議論データのテキストを分析し 12 種類の要素を抽出してアンケート実験でいい議論だと判断された議論データを正解として SVM を用い自動判定への可能性を確認した.

しかし, 先行研究では評価対象が議論全体であり評価項目に従って評価する際に判定の基準が曖昧だという被験者の意見があった．様々な発言が含まれている議論を一つの評価対象にするには補完が必要とされた.

3.3 議論の妥当性モデル

本研究では議論の妥当性を判断するために評価対象を参加者の発言と議論の両方とする．参加者の発言の評価から議論の妥当性を判定し, 議論を評価した結果との比較と検証を行う。その流れは次のようである．まず, 参加者の発言を人が評価し, その評価から参加者の議論スキルを評価値として算出する．そして議論の参加者が持つ議論スキルの評価値から議論の妥当性を表す評価値を導出する．導出された評価値は議論全体を評価した評価値を用いて検証を行う.(図 1)

この提案モデルは「良い参加者が多い議論は良い議論である」という発想から始める．この仮説は様々な研究から前提として使われているが, 本研究では実験を通じてその仮説を検証する.

3.4 議論の特徴

先行研究で用いた論証能力とコミュニケーション能力を議論においてスキルとして定義した．それを本稿では議論スキルと呼ぶ．また効率的結論形成に関しても評価するため「議論を仕

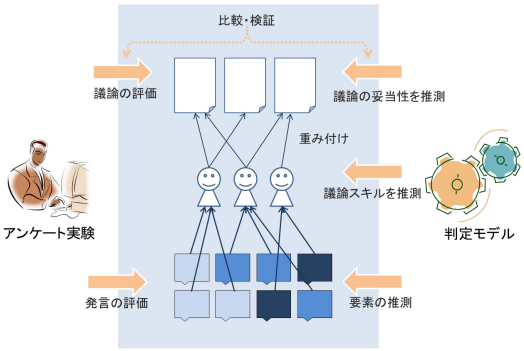

図 1: 議論の妥当性の判定

切る能力」として評価を行う．大坊らは会話コミュニケーションに関する研究 [6] から討論条件では主導的討論者の役割を情報や意見を述べ, 他者への質問や指示による役掛けが多いと主導的討論者の特徴を確認した．この主導的討論者の特徴を本稿では 3 つ目の議論スキル「議論を仕切る能力」として提案する．そこで, 提案した 3 つの議論スキルから 3 つの質問項目と総合的点数の質問項目の総 4 項目の質問を構成した．(表 1)

表 1: 議論の特徴項目とスキル

	議論スキル	質問項目
特徴項目 1	論証能力	根拠をもって論理的に説得をする．
特徴項目 2	コミュニケーション能力	相手を尊重して相手の意見や立場を理解する．
特徴項目 3	議論を仕切る能力	議論の流れを考慮しながら結果を出せるように努力する．
総合判定		総合的点数

4. アンケート実験

本章では前節で提案した評価項目と総合評価, 発言と議論全体の関係性を確認するため発言単位と議論単位で人間がつける評価を行った．

4.1 実験概要

Wikipedia のノートページから収集した議論データは政治・社会 13 件, 歴史 9 件, 文化 5 件, アニメ関連 6 件, その他 8 件総計 41 件であった．複数の議論に参加している場合も含め 41 件の議論データから 461 発言を抽出した．実験対象者の構成は大学生 5 人 (理系 1, 文系 4), 大学院生 (理系 2), 社会人 1 人の 8 人, 平均年齢は 23.5 歳である．被験者に発言単位と議論単位のテキストを配布し, 4 つの質問項目を 7 件法として回答するようにアンケート実験を行った．

4.2 発言単位の評価

前節で定義した質問項目と総合的点数による発言単位の評価を被験者に行わせた．461 件の発言に対してアンケート実験を行った．実験では発言をランダムにして発言の前後が分からないように行った．被験者の評価から平均値を取り 3 つの評価項目と総合評価との関係性を相関係数から確認した．その結果と回帰分析による貢献度を表 2 に示す．

表 2: 発言単位と議論単位の評価

	発言単位の評価		議論単位の評価	
	相関係数	貢献度	相関係数	貢献度
特徴項目 1	0.961652	0.699696	0.900783	0.607886
特徴項目 2	0.422276	0.207318	0.263817	0.204201
特徴項目 3	0.172863	0.101384	0.334604	0.191909

4.3 議論単位の評価

議論単位の評価では議論全体に対しての評価を被験者に行わせた．総合的点数と 3 つの質問項目との相関係数と回帰分析による貢献度を算出した．(表 2)

4.4 考察

アンケート実験を通じて発言単位の評価と議論単位の評価を取った．人による評価では論証能力を主に評価することが分かった．評価項目 3 は議論を仕切る能力を問う評価項目のため, 議論の流れが判断できない単一発言を対象にした実験ではその貢献度が低く見られた．議論単位での評価では発言単位での評価より約 2 倍貢献していることが分かった．これは被験者達にとって先行発言と後続発言の意味が繋がる隣接ペア [7] として認識されたと思われる．

また, 発言単位の評価で評価ができないと判断された発言は 33 件であった．評価できない発言は短文や「(体裁なおし, 文献追加), (文修正)」などの連絡事項などで議論として意味を持たない文章が多かった．平均文章数は 1.187 個, 平均 59.43byte であった．

5. 妥当性モデルの検証

本章ではアンケート実験の結果から議論参加者の評価を行い, 参加者の評価値から議論の妥当性を判定する手法とその検証に関して述べる．

5.1 議論スキル

論証能力, コミュニケーション能力, 仕切る能力を議論参加者の議論スキルとして定義した．今回使われた議論データでの参加者は 144 名のうち判断できない発言を除いた有効参加者は 138 名であった．参加者が持つ議論スキルの評価値は発言の評価値から平均値を取る方法で 3 つのスキルに対して 1 点から 7 点までの評価値として表した．

5.2 議論の妥当性の推測

本稿では参加者の議論スキルから議論の妥当性を評価値として算出する．議論の妥当性を算出するために参加者が持つ議論スキルの評価値に対して参加度を重み付けとして用いる．参加度を表す重み付けとしては「より発言した参加者は議論に大きく影響を与える」と仮定し発言比率を用いた．

5.3 検証・考察

参加者が持つ議論スキルの評価値から計算した議論の妥当性の評価値を検証するため, 議論の全発言に対する評価値を単純に平均値取って比較を行った．特徴項目 1.2.3 の平均値と参加者が持つ議論スキルの評価値から得た議論の妥当性を表す評価値に対して, アンケート実験から得た評価値を順位付けし

て Spearman 順位相関係数を用いて検証した．その比較は表 3 に示す．発言の単純平均より参加者が持つ議論スキルの評価値から得た結果が人による評価により近いと考えられる．複数議論に参加をした参加者や参加者の活動履歴を考慮した評価がその差であり, 発言の単純平均値より優れた結果だと考えられる．

表 3: 提案推測モデルの検証

議論スキル	判定方法	相関係数
論証能力 (特徴項目 1)	提案推測モデル	0.9894934
	単純平均	0.397748
コミュニケーション能力 (特徴項目 2)	提案推測モデル	0.9759849
	単純平均	0.9095684
議論を仕切る能力 (特徴項目 3)	提案推測モデル	0.9868667
	単純平均	0.9566604
総合評価	提案推測モデル	0.9632270
	単純平均	0.8812382

本研究では議論参加者が持つスキルの評価値は参加者の発言から平均値を用いた．評価の高い発言を 1 回した参加者と数回の発言で場合によっては発言の評価が高かった参加者を等しく評価して良いのか疑問が生じた．代表評価値として平均値を用いることが適切なのかその検証を行った．平均値と最頻値, 中央値を代表評価値とした場合の Spearman 順位相関係数を用いて比較を行った．その結果, 表 4 のように平均値が代表評価値としてより適切であり, 参加者 1 人当たりの発言数は 2, 3 回が多かったため最頻値を用いた結果が中央値の場合より高い相関が視られた．

表 4: 代表値の検証

代表値	特徴 1	特徴 2	特徴 3	総合点数
平均値	0.989493	0.975985	0.986867	0.963227
最頻値	0.919699	0.769231	0.845215	0.934896
中央値	0.897936	0.643152	0.820075	0.824577

本研究では議論の妥当性を表す評価値において参加者の発言比率を重み付けとして用いた．その他に文章数比率, Byte 数比率, そしてエントロピー重み付けを用いた結果は表 5 に示す．

表 5: 重み付けの検証

重み付け	特徴 1	特徴 2	特徴 3	総合点数
発言比率	0.989493	0.975985	0.986867	0.963227
文章比率	0.950844	0.952157	0.960037	0.918198
Byte 数比率	0.951031	0.942213	0.968667	0.929643
エントロピー	0.954596	0.470919	0.986867	0.938273

他の重み付けより発言比率からよりいい結果が得られた理由は, 発言量より発言数がより評価されていることと参加者達が発話セグメントの交換から成立する談話としての評価と, そして参加者の積極度が反映されたと思われる．

6. 妥当性の推測モデル

本章ではアンケート実験の結果に基づいて自動的に議論の妥当性を判定するモデルを提案し, その有用性と可能性に関して述べる．

6.1 特徴項目のモデル化

発言データからの要素として表 6 のように文章の特徴要素を抽出した．抽出要素は先行研究 [9] で用いた抽出要素を発言単位の分析に合わせて補完した．

表 6: 文章の特徴要素

1	「から」「ので」の理由節の出現数
2	引用, 出典記号の出現数
3	「れば」「なら」「たら」「ならば」「と」「たらば」の条件節が出現した回数
4	「考える」「思う」「言える」のいずれが語尾に出現した回数
5	「ですか」「ありませんか」「ですかね」「でしょうか」「でしょう」「ますか」「ませんか」などのいずれが質問形の語尾に出現した回数,「なるほど」「〜よね」の語彙的応答系の出現数
6	敬語の割合
7	悪口, 禁止語
8	挨拶の出現数

抽出した要素とアンケート実験から得られた特徴項目の評価値と総合評価の評価値との関係を確認するため回帰分析を行った．まず 8 つの抽出要素が評価項目 1,2,3 と総合評価に与える影響を確認するため相関係数を計算した．その結果は表 7 に示す．

表 7: 抽出要素と項目間の相関係数

要素	特徴 1	特徴 2	特徴 3	総合判定
1	0.5565745	0.0639125	0.1631738	0.5524350
2	0.6596345	0.0457904	0.1052865	0.6239325
3	0.3723432	-0.0533188	0.0892653	0.3497768
4	0.1556887	0.1448566	0.0871259	0.1782472
5	0.0311815	-0.0203003	0.7701857	0.2084326
6	-0.0292784	0.8263307	-0.0687101	0.0569860
7	-0.0193214	-0.2823091	-0.1699335	-0.0842819
8	0.1804785	0.3612971	-0.0753006	-0.1535758

そして, 総合評価と抽出要素から回帰方程式を用いて参加者の議論スキルの評価値を発言から算出し, その評価値から議論の妥当性の判定を行った．

6.2 実験結果

表 8 の結果と発言から抽出した要素を用いて参加者の論証能力, コミュニケーション能力, 議論を仕切る能力, 3 つの議論スキルの評価値を得た．参加者が持つ議論スキルの評価値から議論の妥当性モデルを用いて判定を行った．その結果を表 9 に示す．41 件の議論データの妥当性を判定し, アンケート実験

表 8: 回帰分析の結果

要素	係数	標準誤差
1	1.858188	0.107436
2	1.138512	0.051249
3	2.050212	0.165497
4	0.168797	0.124062
5	0.669115	0.102822
6	0.44226	0.122399
7	0.073053	0.414229
8	-0.3816	0.262307
切片	2.158342	0.116275

で総合評点数が高かった上位 10 件を正解として計算した適合率は 0.7, 順位相関係数は 0.90431267 であった.

表 9: 妥当性推測モデルの実験結果

項目	適合率	順位相関係数
特徴項目 1	0.9	0.95703564
特徴項目 2	0.7	0.94915572
特徴項目 3	0.7	0.93789868
総合判定	0.7	0.90431267

6.3 考察

先行研究の抽出要素を参考にして発言から 8 個の要素を抽出した．論証能力を問う特徴項目 1 では理由節, 引用・出典記号, 条件節の出現頻度が大きく影響を与えていることが分かった．この 3 つの要素は Toulmin モデル [8] の構成要素でもあって論証能力を評価する適切な要素だと考えられる．

コミュニケーションを問う特徴項目 2 では敬語の割合が強く貢献していることが分かった．人によるアンケート実験でも言葉の丁寧さが評価に大きく影響したと被験者の意見があった．

議論を仕切る能力を問う特徴項目 3 では疑問形の語尾や「なるほど」「〜よね」の語彙的応答系が大きい影響を与えていることが分かった．結論を出すため議論を仕切る手段として相手の発言を誘導する行動が評価されたと思われる．

しかし, 先行研究から注目をした悪口, 禁止語にはすべての特徴項目に対して関係性が低いと視られ, 特徴を表す要素をより多く取り組むことが今後の課題となった．

7. 考察と今後の課題

今回 Wikipedia のノートページから 41 件の議論データを集め, 461 件の発言を実験に用いた．アンケート実験を考慮してデータを収集したため参加者が持つ議論スキルと議論の妥当性モデルの有用性を判断するに当たりデータの数が少ない．今回の実験を基盤として今後大量の議論データを分析して妥当性を判断するモデル適用して行きたい．

アンケート実験から人は論証能力に関する部分と共にコミュニケーション能力, 議論を仕切る能力に関しても注目していることが分かった．いい議論に発展させるためには議論参加者にいわゆる社会性が要求されていると考えられる．

発言単位の評価での議論を仕切る能力を問う評価項目 3 の貢献度が議論単位の評価では約 2 倍大きくなった．前後関係の判断ができない発言単位の評価では議論を仕切る能力を評価することが不適切であった．今後こういった議論の流れから発見される特徴の定義に関しても改善して行きたい．

8. おわりに

本稿では, 議論の妥当性を判断するために 3 つの特徴を定義して特徴項目を決めた．参加者の発言と議論全体に対する分析から参加者が持つ議論スキルの評価を行った．参加者が持つ議論スキルの評価値から議論の妥当性を表す評価値を導出して議論参加者の能力が議論全体に影響を与えることを確認した．またテキストから特徴要素を抽出し, 提案した特徴項目との関係性を用いて参加者と議論の妥当性を判定する手法への可能性も確認した．議論参加者のスキルから議論の妥当性を判定することによって議論における充実度の判定や結論形成のための支援を議論参加者側から調整することが可能になったと思われる。今後の課題としては厳密な分析を通じてより多様な要素を抽出し, 議論参加者が持つスキルの評価値と傾向をより正確に算出できるように改善して行きたい．また, この判定モデルから効率的な議論に発展させるための支援をする枠組みを提案していきたい．

参考文献

[1] 鈴木 優，吉川 正俊, Wikipedia におけるキーパーソン抽出による信頼度算出精度および速度の改善，情報処理学会論文誌: データベース，Vol.3, No. 3 (TOD47), pp. 20 - 32, 2010 年 9 月

[2] 鈴木 優，金本 径卓，川越恭二, Wikipedia の編集履歴を用いた記事の信頼性導出，人工知能学会第 20 回セマンティックウェブとオントロジー研究会，2009 年 2 月

[3] 桜井茂明，折原良平，掲示板サイト分析における重要議論抽出と特徴表現抽出，知能と情報，Vol.19，No.1，pp.13-21，(2007).

[4] 松村真宏，加藤優，大澤幸生，石塚満, 議論構造の可視化による論点の発見と理解 , 日本ファジィ学会誌，Vol.15, No. 5，pp. 554-564 (2003).

[5] Beer, Francis A., Validities: A Political Science Perspective, Social Epistemology 7, 1 pp. 85-105 (1993).

[6] 大坊郁夫，社会的スキル向上を目指す対人コミュニケーション, ナカニシヤ出版, pp202-117, (2005).

[7] 坊農真弓, 高梨克也，多人数インタラクションの分析手法, オーム社, pp.82-94, (2009).

[8] Toulmin, S.E., The Uses of Argument, Cambridge University Press. (1958).

[9] 朱成敏, 武田英明, Wikipedia における編集者間議論ページの分析とそのモデル化, 第 21 回 Web インテリジェンスとインタラクション研究会電子情報通信学会 2011 年 9 月.