RK-001

CGM レビュー評価の分布と炎上状態の関係
Relation between the Distribution of CGM Review and the Flaming Discussion

飯尾　淳 †
Jun Iio

1. 本研究の背景

ブログや SNS，あるいは twitter などのいわゆるマイクロブログの流行により，一般消費者が簡単に情報を発信することができるようになった．消費者が簡単に情報コンテンツを公開することができるメディアを総称し，CGM（Consumer Generated Media）と呼ぶ．インフラが整い，初心者でも簡単に情報を発信できるようになったため，現在は CGM が大流行，一種の社会現象ともなっている．

また，オンラインショッピングのシステムでは，販売している商品にユーザーがレビューを寄せることが可能となっているシステムも多い．そのようなシステムでは，商品を購入したユーザーが何段階かの点数をつけることができるようになっている．

しかし一方で，このようなレビューシステムを悪用した「ステルスマーケティング」と呼ばれる情報操作も現れている．ステルスマーケティングとは，CGM によるレビュー評価を悪用し，あたかも消費者の口コミで高評価が拡がっているように情報操作を行うマーケティング手法である．口コミによるマーケティング手法をバイラルマーケティングとも呼ぶが，ステルスマーケティングは，ある種の信頼で成り立っているバイラルマーケティングや CGM の効果を悪用した手法と指摘できる．このように，CGM 全盛の現代においては，一般消費者から寄せられたレビューの質が問われている時代となった．

2. ユーザーレビューの問題

前述した背景のもとで，CGM の問題点を指摘したい．CGM ではしばしば，炎上という状況が発生する．炎上，すなわち議論が白熱して刺激的なコメントの応酬が続く状況や，あるいは過度に感情的になり議論といえない状況や社会的にほとんど意味のない状況に陥るケースが発生する．炎上状態はネットの初期から電子掲示板等でみられていたが，CGM 全盛の今日，普遍的にみられる現象となった．オンラインコマースのシステムでも，いわゆる「やらせ問題」が発覚して炎上したケースが報じられている [1].

システムの健全性を維持するためには，炎上状態は早期に鎮火，解消することが望ましい．そのため，炎上を機械的に発見する方法が望まれている．これまで炎上の早期発見に関する研究はいくつか報告されている [2, 3] が，できるだけ客観的な判定を可能とするために，テキストマイニング等によらず，レビューの内容に踏み込まず炎上を判定する方法を探求したい．

本研究では，ユーザレビューによる評価値の分布に着目することにより，炎上状態を機械的に発見することを目的として，ユーザレビュー評価値の分布とユーザレビュー自体の評価に関する相関を調べた．次節で説明するように，ユーザレビュー自体の評価が持つ特徴を，炎上状態に関連付けて考察を加えた．

3. レビューコメントの分析

多くのユーザレビューシステムでは，一般消費者が投稿したそれぞれのレビューコメントや評価結果に対しても，他のユーザが評価を加えることができる．例えば代表的なオンラインショッピングサイトである Amazon では，それぞれのレビューに対し，「このレビューは参考になりましたか」という問いが用意されており，それに対して「はい」もしくは「いいえ」という評価を加えることができる．これを利用すると，ユーザレビュー k に対する投票の総数を V_k^{total},「参考になる（はい）」と答えた投票数を V_k^{agree} としたとき，そのレビューの有用度 w_k を，以下の式で定義[*1]することができる．

$$w_k = \begin{cases} V_k^{\text{agree}} / V_k^{\text{total}} \\ 0, \text{ if } V_k^{\text{total}} \text{ is 0.} \end{cases}$$

図 1 から図 3 に示す 3 つのグラフは，このように定義した有用度を x 軸に，レビューコメントの長さを y 軸にとり，Amazon（日本）で販売されている書籍からレビューが多く寄せられている作品を抽出してレビューコメントの分布を示した[*2]ものである．

作品 A は，全体として評価が高く，星 3 つから星 5 つまでの高評価が数多く分布している．それに対し，レビューの評価は全体的に均一に散らばっているという傾向がみられる．作品 B はそれよりも評価がやや低めになっている．レビューの評価は，低い評価ほど有用度が高くなる傾向をみせる．また作品 C は著者が意外な人物ということで話題をさらった作品である．内容以外で評価されたのではという理由からレビューコメントは荒れ，炎上の傾向をみせた．評価は高い評価と低い評価に割れ，評価スコアと有用度には相反する関係をみてとることができる．

このように，レビュー対象の性質に応じてコメントの評価も変わる．そこで，レビューコメントの分布から，コメントの状況を推定する，すなわちレビューに対する評価から炎上を判定できないかという方法の検討を試みた．

なおレビューが荒れていることの指標として，以下の指標を用いる．

レビューによる評価スコアを単純平均した $\bar{S}$ に対し，各レビュー k の寄与度 $w_k/\bar{w}$ を有用度 w_k から求める．

† （株）三菱総合研究所, MRI

[*1] V_k^{total} が 0 のとき w_k の値を 0 から 1 の間でどの値にするべきかについては，検討の余地が残る．

[*2] y 軸は対数目盛となっている点に注意されたい．なおレビューの長さと有用度の直接的な関係はなく，y 軸は単に散らばり具合を分かりやすく示すためだけに利用されている．

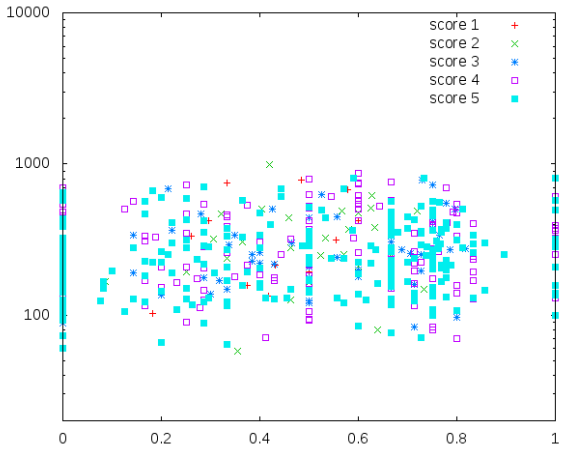

図 1: 作品 A（直木賞受賞作）

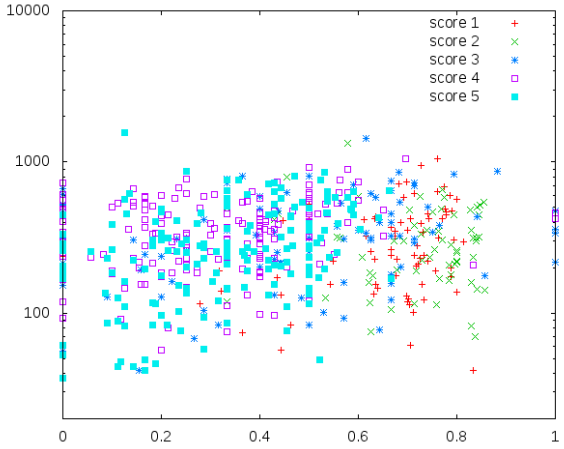

図 2: 作品 B（本屋対象受賞作）

その上で，寄与度を重みとして用いた加重平均を以下のとおり計算する．

$$\bar{S}_w = \frac{1}{n}\sum_k \frac{w_k}{\bar{w}} S_k = \frac{1}{n}\sum_k \frac{w_k}{\frac{1}{n}\sum_i w_i} S_k = \frac{1}{\sum_i w_i}\sum_k w_k S_k$$

この加重平均はユーザレビュー自体の評価で左右されるものであり，ユーザレビューに対する評価に偏りがみられると，単純平均の値とは異なる値になる．したがって，単純平均の値と比べて差が大きいレビューは，レビュー自体に何らかのバイアスがかかっている状態であると推測することができる．

4. 実験

本節で，実験の方法と結果について述べる．実験は Amazon.co.jp で公開されている年間ベストセラーを対象に実施し，前述した平均のずれとレビュー評価の分布に関する相関を確認した．

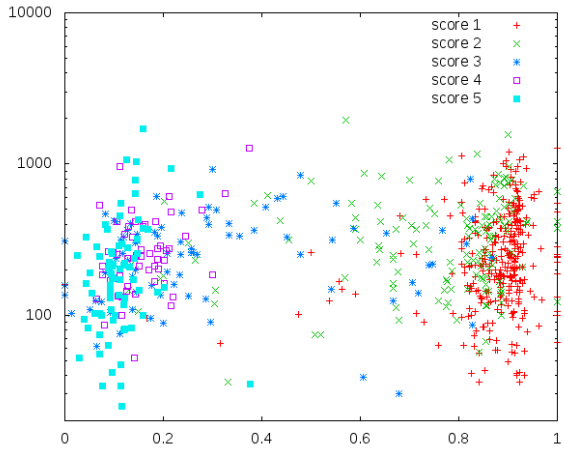

図 3: 作品 C（ポプラ社小説大賞受賞作）

4.1. 実験方法

Amazon.co.jp には，年間ベストセラーのアーカイブが公開されている．同アーカイブには 2000 年から現在までのデータが蓄積されているが，2000 年は途中からのデータであり，また 2012 年も途中までのデータのため今回の対象からは除外することにし，2001 年から 2011 年までの年間ベストセラーとして列挙されているアイテム（書籍）を実験の対象とした．

これらのアイテムを対象にレビューのデータを取得する．レビューの評価としてそれぞれのレビューでユーザにより評価されている星の数を 1 から 5 までのスコアとして用い，さらに各レビューに対して寄せられた有用度のデータも取得した．

なお全てのデータは用いず，レビューが多く寄せられているもの（具体的にはレビュー数が 100 以上のもの[*3]）のみを対象とした．統計的な観点に加えて，そもそもレビューが少ないものは炎上の余地がないため，除外してよいと考える．

4.2. 実験結果 1: ユーザによるレビューの補正

Amazon.co.jp の年間ベストセラーは，同サイトで販売されたアイテムのうち販売に関して上位 100 位までのアイテムがリストアップされている．したがって候補としては 1,100 アイテムが対象として考えられたが，数年にわたり上位 100 位中にランクインしている書籍（重複）があること，ごく僅かに書籍以外の物品が含まれていること，また既に販売が終了しており詳しい情報を取得できないものが数点含まれていたこと，以上を鑑み，それらを排除した結果，結果として 882 個のアイテムが対象となった．

さらに各アイテムに寄せられたレビュー数が 100 を超えないものを除外すると，204 個のアイテムが残った．これらを対象に，先に述べた加重平均と単純平均の差を計算した．レビュースコアに関する加重平均値

[*3] この 100 という数値選択の根拠はなく，実際にレビュー数が何件以上あれば十分な検証ができるかの議論は今後の課題である．

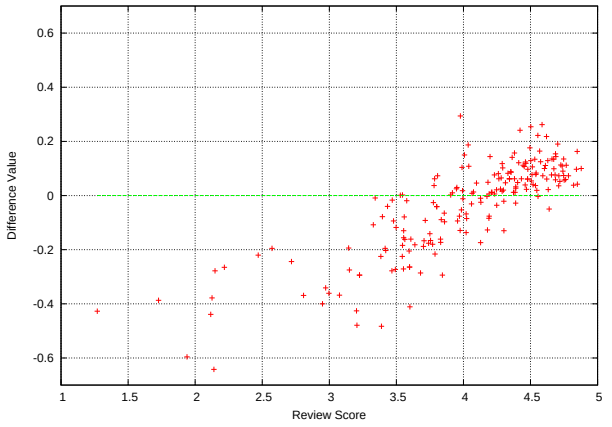

図 4: レビューのスコアと，単純平均 – 加重平均の差

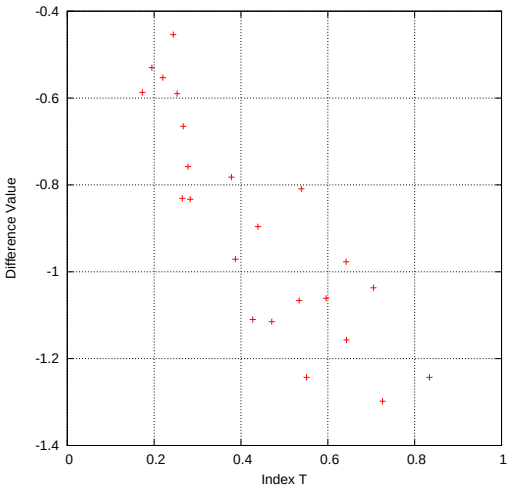

図 5: 指標 T_k と補正値の関係

と補正値（単純平均 – 加重平均）の関係を図 4 に示す．グラフから，ユーザによるレビューの補正が行われると，低い評価はより低く，高い評価はより高くなるという傾向を確認することができる．

また，レビューの補正で評価が大きく変動したアイテムの例を表 1 に示す．表 1 には，補正値の絶対値で順序付けたときに 1 位から 10 位までの例を示した．ほとんどのアイテムで，レビューのスコアは低い方向に補正されている．

4.3. 実験結果 2: レビューの分布と補正値の関係

表 1 に示された多くのアイテムでは，低い評価をさらに押し下げる方向に補正されている傾向がみられる．そこで，レビューの分布に関して正規分布からの外れ度合いとして新たな指標を導入する．具体的には，アイテム k に関するそれぞれ 1 から 5 までの評価 i に対し，レビューの分布が正規分布を仮定したときの割合 $t_0(i)$ と実際の割合 $t_k(i)$ の差の絶対値を考え，その総和としての指標 T_k を以下で定義する．

$$T_k = \sum_{i=1}^{5} |t_0(i) - t_k(i)|$$

表 1: レビュー補正で評価が大きく変動した書籍の例

タイトル	レビュー数	補正値
働かないで年収 5160 万円稼ぐ方法	144	-1.400
脳を活かす勉強法	132	-1.012
新しい歴史教科書―市販本	138	0.826
頭の回転が 50 倍速くなる脳の作り方 ～「クリティカルエイジ」を克服する加速勉強法～	110	-0.835
ＫＡＧＥＲＯＵ	750	-0.726
なぜ、社長のベンツは 4 ドアなのか? 誰も教えてくれなかった!裏会計学	156	-0.642
世界の中心で、愛をさけぶ	1009	-0.596
無理なく続けられる 年収 10 倍アップ勉強法	195	-0.483
お金は銀行に預けるな 金融リテラシーの基本と実践 (光文社新書)	200	-0.479
下流社会 新たな階層集団の出現 (光文社新書)	357	-0.439

この T_k は，アイテム k に関するユーザレビューの分布形態を表す指数として捉えることができる．この指数と炎上の関係をみるために，先ほど求めた補正値 $\bar{S} - \bar{S}_w$ と T_k の相関を求めた．

なお，低い評価が多いものが炎上しやすい傾向があることから，評価の分布で低評価のほうが多いもの，すなわち評価の分布で回帰直線を引いたときに，係数が負になるものを抽出した．ただし先ほど対象とした 204 個のアイテムからこの条件に合致するものを抽出すると数が極端に少なくなってしまうため，レビュー数の条件を 10 個以上に緩めた．

指標 T_k と補正値のグラフを図 5 に，またそれぞれの値を表 2 に示す．Spearman の順位相関係数を求めたところ，相関値 $\rho = -0.876$ と負の高い相関を得た．

以上から，低い評価のレビューが正規分布から外れるような状況で多数寄せられている状況が発生しているケースでは，レビューに対する評価も荒れる，すなわち炎上が発生している可能性が高いということが推測される．そもそもレビューが多く寄せられるという状況はそのアイテムが数多く市場に出ているということであり，低い評価が数多く投稿されるというケースは一種の特異な事態である．それを踏まえると本実験

表 2: 指標 T_k と補正値の関係

ISBN	指標	補正	ISBN	指標	補正
459112245X	0.73	-1.30	4062569671	0.55	-1.24
4391631008	0.83	-1.24	4820300547	0.64	-1.16
4846307069	0.47	-1.12	4140880368	0.43	-1.11
490362093X	0.53	-1.07	4093860726	0.60	-1.06
4871772349	0.71	-1.04	4894512262	0.64	-0.98
409386280X	0.39	-0.97	4334033210	0.44	-0.90
4757526482	0.28	-0.83	4106101378	0.27	-0.83
4894512963	0.54	-0.81	408781372X	0.38	-0.78
4569635458	0.28	-0.76	4915512703	0.27	-0.67
4087805026	0.25	-0.59	4163706208	0.17	-0.59
4569657052	0.22	-0.55	4106100037	0.20	-0.53
4087746836	0.24	-0.45			

の結果はごく自然な結論であると考えられる．

5. 関連研究

ユーザによるレビューや評価が一般化して以来，CGM のレビューに関する研究は数多く行われている．

Wang ら [4] はレビューの信頼性と時間で劣化するランキングシステムを提案した． また Li ら [5] は Web サービスを比較するためのレビューランキングシステムを提案している．Ghose ら [6] はまた別のランキングアルゴリズムの提案を行った．このように，レビューを評価しランキングを行う研究は多数 [7, 8, 9, 10] 実施されている．

レビューを自動で分類しようという試み [11, 12, 13, 14, 15, 16, 17] も多い．例えば Chua ら [15] はレビューアに着目し，機械学習の方法を用いてレビューアを順位付けする方法を提案した．

またスパム的なレビューを自動で検出する方法も検討されている [18, 19]．スパムだけでなく，品質の低いレビューを検出しようとする試み [20] もある．Hu ら [21] はレビューの多くは非対称の分布を示すことが多いと指摘している．本論文で示した正規分布から外れるものに着目するというアイデアに似た発想である．

6. まとめと今後の課題

本研究では，ユーザレビューによる評価スコアの単純平均と，それぞれのレビュー自体を他のユーザが評価した値を用いた加重平均を考え，それらの差分による補正値をレビューが荒れている状況を示す数値として扱った．また，ユーザレビューの炎上状態を発見するための指標として，低い評価が多く寄せられている状態を示す指標を導入した．

実験により，これらの値には高い相関があることを導き，レビューコメントの中身に触れずともレビューが荒れている状況を発見するために利用できる可能性を示した．結果は極めて自然なものであるが，定量的に評価できたことは，炎上状態を自動判定できる可能性を示唆するものである．

今回実施した実験では，テキストマイニング等によるレビューの評価結果と突き合わせた検証は実施していない．したがって本当に炎上しているかどうかの客観的な評価は未検証である．実際に「問題のありそうな」アイテムのレビューデータを目視で確認すると炎上状態を確かめることができるが，それらを機械的に検証し本手法の有効性を明示することが，今後の課題として残されている．

参考文献

[1] 滝田, “実は”やらせ”だったという落とし穴... ネットのクチコミ情報にご用心!,” 週刊朝日 114(27), pp. 36–38, 2009.

[2] 橋本, 白田, “ソーシャルコンピューティング可視化サービスの検討,” 情報処理学会研究報告. データベース・システム研究会報告 2009-DBS-149(23) pp. 1–7, 2009.

[3] 亀井, 和田, 大西, “ネット上で勃発する風評被害 — 監視方法と「炎上」の早期発見と初期対応策,” ビジネス法務 11(9), pp. 29–33, 2011.

[4] B.C. Wang, W.Y. Zhu, and L.J Chen, *Improving Amazon-like Review Systems by Considering the Credibility and Time-Decay of Public Reviews*, Informatica, Vol. 35, No. 4, pp. 463–472, 2011.

[5] H.H. Li, X.Y. Du, and X. Tian, *A Review-Based Reputation Evaluation Approach for Web Services*, Journal of Computer Science and Technology, Vol. 24, No. 5, pp. 893–900, 2009.

[6] A. Ghose and P.G. Ipeirotis, *Designing Novel Review Ranking Systems: Predicting the Usefulness and Impact of Reviews*, Proceedings of the 9th International Conference on Electronic Commerce, pp. 303–310, 2007.

[7] C.D.N. Mizil, G. Kossinets, J. Kleinberg, and L. Lee, *How Opinions are Received by Online Communities: A Case Study on Amazon.com Helpfulness Votes*, Proceedings of the International Conference on the World Wide Web, 2009.

[8] G. Ganu, N. Elhadad, and A. Marian, *Beyond the Stars: Improving Rating Predictions using Review Text Content*, Proceedings of the 12th International Workshop on the Web and Databases, Province, RI, June 2009.

[9] P. Chen, S. Dhanasobhon, and M.D. Smith, *All Reviews are Not Created Equal: The Disaggregate Impact of Reviews and Reviewers at Amazon.com*, SSRN eLibrary, 2008.

[10] W.H. Davidson, M. McLeod, C. Klerkx, and M. Wojcik, *A Method for Measuring Helpfulness in Online Peer Review*, Proceedings of the 18th ACM International Conference on Design of Communication, pp. 115–121, São Paulo, Brazil, 2010.

[11] H. Cui,V. Mittal, and M. Datar, *Comparative Experiments on Sentiment Classification for Online Product Reviews*, Proceedings of the 21st National Conference on Artifical Intelligence, 2006.

[12] M.P. O’Mahony, P. Cunningham, and B. Smyth, *An Assessment of Machine Learning Techniques for Review Recommendation*, Proceedings of Artificial Intelligence and Cognitive Science: twentyth Irish Conference, AICS 2009, Dublin, Ireland, August 2009.

[13] M.P. O’Mahony and B. Smyth, *A Classification-based Review Recommender*, Journal of Knowledge-Based System, Vol. 23, No. 4, pp. 323-329, 2010.

[14] S.M.Kim, P.Pantel, T.Chklovski, and M.Pennacchiotti, *Automatically Assessing Review Helpfulness*, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 423–430, Sydney, July 2006.

[15] F.C.T. Chua, *Summarizing Amazon Reviews using Hierarchical Clustering*, Technical Reports, available at `http://www.mysmu.edu/phdis2009/freddy.chua.2009/`

[16] Z. Zhang, *Weighing Stars: Aggregating Online Product Reviews for Intelligent E-commerce Application*, Intelligent Systems, IEEE, Vol. 23, Issue:5, pp. 42–49, September 2008.

[17] L. Qu, G. Ifrim, and G. Weikum, *The Bag-of-Opinions Method for Review Rating Prediction from Sparse Text Patterns*, Proceedings of the 23rd International Conference on Computational Linguistics, Stroudsburg, PA, 2010.

[18] N. Jindal and B. Liu, *Review Spam Detection*, Proceedings of the International Conference on the World Wide Web, Poster Paper, 2007.

[19] E.P. Lim, V.A. Nguyen, N. Jindal, B. Liu, and H.W. Lauw, *Detecting Product Review Spammers using Rating Behaviors*, Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 2010.

[20] J. Liu, Y. Cao, C.Y. Lin, Y. Huang, and M. Zhou, *Low-Quality Product Review Detection in Opinion Summarization*, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 334–342, Praque, June 2007.

[21] N. Hu, P.A. Pavlow, and J. Zhang, *Can Online Word-of-Mouth Communication Reveal True Product Quality? Experimental Insights, Econometric Results, and Analytical Modeling*, Proceedings of the 7th ACM Conference on Electronic Commerce, pp. 324–330, 2006.