3P-3

Web書評を用いた書籍情報の分析

中山翔介† 冨山北斗‡ 伊東栄典* 廣川佐千男*

† 九州大学理学部物理学科情報理学コース
‡ 九州大学大学院システム情報科学府 * 九州大学情報基盤センター

1 はじめに

現在 Web には，個人が自分の嗜好に基づいて商品について評価する評判情報が増大している．評価の対象としては，書籍や映画，芸能人，様々な商品，レストランなど様々なものがある．

その理由に，Blog や SNS(Social Networking Site) といった，個人が情報発信することのできる CGM(Consumer Generated Media) の利用者数の急増がある [3]．Blog や SNS では，個人が自分の嗜好に基づいて商品に対する評判を記述することが多い。これはアフィリエイトも関係している．

また，Web 上の商品販売サイトにおける，利用者レビューも増大している．販売サイトでは販売促進のために，商品に対する利用者からの評価 (review) を記載することが多い．たとえば Amazon.com では書籍に対する書評を記載している，kakaku.com では PC 等の電子機器についての評価を記載している．

我々は個人的な嗜好に基づいた評判情報の分析を行っている [2, 1]．また，個人的な嗜好に基づく，アイテム (商品) 推薦についての研究も行っている [1]．本研究では，具体的な対象として書籍を選び，個人書評を利用した，書籍推薦システムの構築を目指している．本稿では，システム開発のための，書評 Wiki における書評情報の分析について述べる．

2 Web書評

2.1 「書評 Wiki」

「書評 Wiki」[4] は,「Mystery Laboratory」というサイトの管理人 matsuo 氏が管理運営している Wiki サイトである．サイトは PukiWiki システムを利用しており，誰でも自由に書評を書き加えることができる．ここには書評のみが蓄積されているため，ノイズを含まない書評を閲覧可能である．

2.2 書籍販売サイトの書評

Amazon.co.jp に代表されるように，書籍販売サイトでは，書籍に関するレビューを公開している．こうしたサイトでは，利用者からのレビュー投稿を DB を用いて管理している場合が多いため，レビューの構造が特定できるため，系統的な利用が可能である．ただし，レビュー情報を分析するためには，その DB 内の情報を入手する必要がある．

2.3 Blog や SNS 内の書評

Blog や SNS では，個人が私的な嗜好に基づいてある特定の商品に対する評判を記述することが多い。書籍は評価を書きやすい対象であるためか，書籍への感想の記述は多い．

Blog や SNS から書評を収集には問題がある．まず，書籍関連情報の特定が困難である．Blog や SNS 内では雑多な記述が多く，書籍に関連する記事かの判定が必要になる．また，作家と書籍タイトルの特定が問題になる．

さらに，収集量も問題にある．Blog は公開されていることが多いが，SNS は基本的に閉鎖されているため，大量の情報収集は困難である．

2.4 Web 掲示板内の書評

「2ch」に代表される Web 掲示板の利用者も増加している．Web 掲示板内には多数の書評が有るものの，書込み数の膨大さとノイズ率の高さから，掲示板からの書評特定は困難である．

3 傾向分析

書評の分析の手始めとして，入手が容易な「書評 Wiki」から書評を入手した．本節ではその分析につい

て述べる．

3.1 傾向分析

研究室で作成しているクローラー [5] を用い，200 年 10 月 23 日から 25 日の三日間で書評の収集を行った．収集したファイルは，31,169 個で，総ファイルサイズは 965M である．「書評 Wiki」において，書評の記述がある作家数は 2,923 人である．また，識別できるレビュアー数は 4,537 人であるただし，「書評 Wiki」では，一人のレビュアーが別名で書評を記述することが可能であるため，現実のレビュアー数とは異なる可能性がある．

書評ファイル数	31,169	個
作家数	2,923	人
レビュアー数	4,537	人

3.2 レビュアー・作家数の分析

書評件数の多い作家は，人気作家であると言える．そこで，まず一人の作家あたりの書評件数を分析した．紙面の都合から平均書評件数だけを述べる．作家あたりの平均書評件数は 10.363 件であった．

次に各レビュアーが記述した書評の数を分析した．図 1 に，レビュアーごとの書評数をプロット図を示す．X 軸はレビュアー，Y 軸はそのレビュアーの書いた書評数である．どちらも log scale にしている．また，X 軸のレビュアーは，書評数の多い順で並べている．なお，レビュアー一人が記述した書評の平均数は 2.395 人である．

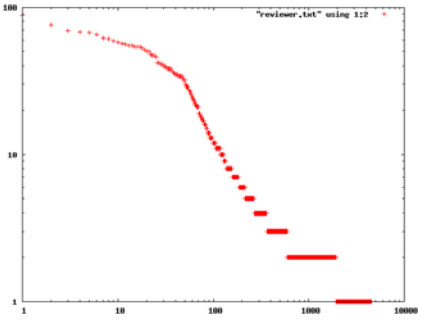

図 1: レビュアーごとの作家数

図 1 は，Zipf の法則とは異なる傾向を示している．Zipf の法則では直線の冪分布になるが，図 1 では 3 箇所に折れ曲がりがある線となっている．これは書評特有の現象かもしれないし，収集した書評数が少ないためかもしれない．今後の詳細な分析が必要である．

3.3 関連分析

研究室で作成してる「概念グラフ」[2] を用いて，作家間の関係を分析を行った．概念グラフでは文書群から単語間の関係を抽出する．予め，レビュアー毎に，そのレビュアーが書いた書籍の作家名のリストを作成し，リストを文書，単語を作家名として置き換えて概念グラフを構築した．

図 2 に，「赤川次郎 OR 村上春樹 OR 村上龍」を入力し，作家間の関係を示す．

図 2: 作家間の関係 (赤川次郎，村上春樹，村上龍)

4 おわりに

本稿では，Web 上に存在する書評（書籍へのレビュー）を用いた書籍情報の分析についての考察を行った．また，「書評 Wiki」にある書評の分析を行った．今後は，更なる詳細な分析を行う予定である．また，Amazon などの書籍販売サイトからの書評収集も行う予定である．最後に，評判情報に基づく推薦システムを開発する予定である．

参考文献

[1] Yufeng Dou, Eisuke Itoh, Sachio Hirokawa, Daisuke Ikeda : “An Approach to Analyzing Correlation between Songs/Artists Using iTMS Playlists”, Proc. IAWTIC’2005, Nov., 2005.

[2] 廣川佐千男, 下司義寛, 和多大樹 : “文書群からの概念グラフの構成”, 情報処理学会第 169 回自然言語処理研究会, pp.79〜84, 2005.

[3] http://japan.internet.com/wmnews/20051128/5.html , 2005.

[4] “書評 Wiki”, http://mystery.parfait.ne.jp/wiki/pukiwiki.php, 2005.

[5] Y.Matsunaga, S. Yamada, E. Ito and S. Hirokawa : “A Web Syllabus Crawler and Its Efficiency Evaluation”, Proc. of ISEE2003, pp.565–568, 2003.