ラフ集合理論を用いたマンガ推薦システムの性能評価

2013SE246 山口健吾　2014SC050 村井諒次
指導教員：河野浩之

1　はじめに

スマートフォンの普及で気軽にインターネットが利用できるようになった現代において, 多くの作家は自分の作品を世の中に公開する手段が増加してきている．Twitter などの SNS や, comico やマンガワンといったスマートフォン向けマンガアプリ, pixiv といったイラスト投稿サイトなど数多く存在する．それらで公開した作品に人気が出て単行本化することも少なくない．他にもアニメやライトノベルのメディアミックスでマンガ化することもある．そういったことから, 2016 年に出版された新刊コミックの点数は出版指標年報 2017 年版 [1] によれば 12,500 点を上回る．このように日々増加していくコミックの中からユーザの興味や好みに沿った作品を探し出すことは非常に困難である．

本研究では, ウェブスクレイピングとクローリングを用いて作品データベース内のマンガランキングから情報を抽出し, マンガに関する言及や感想などのマンガに関連するテキストを解釈することで間接的にコミック内容を把握し, ラフ集合を用いてユーザの興味に沿った内容の各作品に対する情報アクセスを可能にするアプローチをとることで, コミックの内容情報に基づく情報アクセスを可能にするマンガ推薦システムの構築を行う．

本論文は 5 章で構成されている．2 章では推薦システムに関する先行研究について紹介する．3 章ではラフ集合を用いたマンガ推薦システムの提案を行う．4 章ではスクレイピングの説明や実験結果, ラフ集合のプログラムを示す．5 章では今後の課題を示す．

2　推薦システムの先行研究

本章では, 本研究に対する先行研究について紹介する．2.1 節ではコミックの内容情報に基づいた情報アクセスの手段について, 2.2 節ではラフ集合を用いた推薦手法について, 2.3 節では各論文の比較について述べる．

2.1　コミックの内容情報に基づいた探索的な情報アクセスの手段 [2]

山下らは, コミック作品に関連するレビューなどから情報の抽出を行い, 抽出した情報と類似する作品を提示することを繰り返すことでユーザの嗜好に合致した作品を推薦するシステムの研究を行った．この研究では著者名などの明確とした情報を持たない情報欲求が曖昧なユーザを対称とした推薦システムの構築を行うものである．提案したシステムはユーザの嗜好に合致した作品へのアクセスができたという結果を得ている．

2.2　ラフ集合を用いた感性のモデル化に基づいた推薦手法 [3]

西澤らは, インターネット通販サイトでの閲覧履歴からユーザの好みに合致した商品を推薦するサービスでは本当にユーザの好みに合致した商品を推薦できるとは限らないとして, ラフ集合を用いて各個人の好みに適する洋服をデータベース内から検索する個人適応型 Web サイトを構築し, ユーザが理解できる推薦を行うことを目的としている．この研究では, 個人適応型の情報検索システムを実現するために, 対象に関する特徴を抽出し, かつ, その特徴と人の好みとの関係の度合いを推定している．提案した手法はユーザの嗜好を推測する正解率が大幅に高いことが結果として得られた．

2.3　先行研究の比較

前で述べた論文の比較を行う．[2] の研究では, ユーザの興味に沿ったコミックにアクセスすることができ, [3] の研究ではユーザの嗜好を推測しユーザの好みに沿った洋服を推薦することができた．

表 1　先行研究の比較

先行研究	内容	結果
[2]	レビューなどから情報を抽出し類似する作品を提示	嗜好に合致した作品へのアクセスができた
[3]	ラフ集合を用いて好みに適する洋服を推薦	ユーザの嗜好を推測する正解率が高い

3　ラフ集合理論を用いたマンガ推薦システムの提案

本章では, 本研究のマンガ推薦システムの提案について示す．3.1 節では, 本研究のマンガ推薦システムについて説明する．3.2 節ではウェブスクレイピングについて, 3.3 節ではクローリングについて示す．3.4 節ではデータベースツールについて示す．3.5 節ではマンガデータベースサイトについてと利用する理由について説明する．3.6 節ではラフ集合理論について説明する．

3.1　ラフ集合を用いたマンガ推薦システム

ラフ集合理論を用いたマンガ推薦システムの概要図を図 1 に示す．初めに, マンガのレビューが掲載されているサイトからスクレイピングとクローリングによりそれぞれのマンガに対するレビューを切り出し, 切り出したレビューを

Mecab を用いて形態素解析を行う．形態素解析によって抽出した面白いや感動, 恋愛などといった条件属性となる属性情報を整理しマンガデータベースを作成する．作成したマンガデータベースはラフ集合で用いるためにマンガに対してそれぞれの属性情報の有無を記述して作成する．その際, マンガのレビューが掲載されているサイトのマンガに対する点数を用いて作品に対して好きか嫌いかの決定属性を入力しまとめる．作成したマンガデータベースをラフ集合で用いる決定表として if-then ルールの抽出を行う．ラフ集合によって抽出された if-then ルールをまとめたものをルールデータベースとする．作成したルールデータベースとマンガの属性情報の有無が記述されている表を用いて好みの推測を行う．

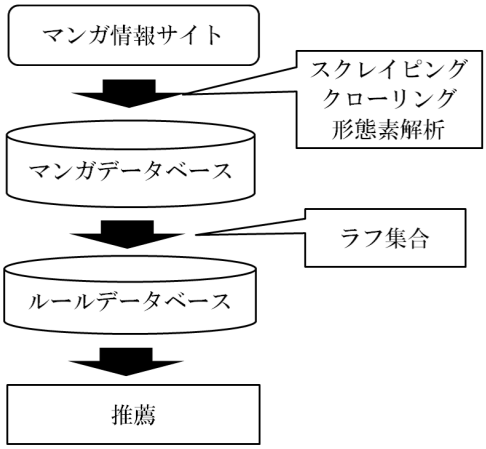

図 1　システム概要図

3.2　ウェブスクレイピング

本研究ではデータを収集するにあたってウェブスクレイピングを用いる．ウェブスクレイピングとはプログラム上でウェブサイトにアクセスし, ウェブサイト上の情報から必要な情報だけをを抽出するコンピュータソフトウェア技術である．例えば料理サイトから材料だけを切り取って表示したり, まとめサイトのタイトルだけを切り取ったりすることができる．ただし, 短い間隔で大量に行うと DOS 攻撃になり相手のサーバをダウンさせてしまう恐れがある．

ウェブスクレイピングは Ruby や PHP, Python など様々な言語で行うことができる．Nokogiri は Ruby のスクレイピングライブラリで Xpath や CSS セレクタを使った要素の抽出を行うことができる．Goutte は PHP のライブラリで, 他のモジュールを使わずに URL へのアクセスが可能である．BeautifulSoup は Python のライブラリで, HTML と XML のパーサーでパースツリーを作成することができる．本研究では Ruby 言語の Open-uri というモジュールと Nokogiri という構文解析器を用いる．Open-uri は URL から HTML を取得し, 内容を文字列として返すことが可能であり, Nokogiri は取得した HTML から CSS 記法で欲しい情報を選び切り出しを行う．

表 2　スクレイピングの比較

ライブラリ名	言語	特徴
Nokogiri	Ruby	記述が簡単
Goutte	PHP	URL へのアクセス可能
BeautifulSoup	Python	パースツリーを作成可能

3.3　クローリング

クローリングとはプログラムがインターネット上の Web サイトの HTML に記載されているリンクを辿り Web サイトを巡回し Web ページ上の情報を複製・保存を行うことである．クローリングを用いることで多くのページから自動で情報を収集することが可能になる．また, クローリングで情報を収集しスクレイピングでデータの解析を行い, データを保存することをクローラーと呼ぶ.

ウェブスクレイピングと同様にクローラーも様々な言語で行うことが可能である．本研究ではウェブスクレイピングを Ruby 言語で行うため, クローリングも Ruby 言語を用いる．Ruby 言語では Anemone というクローラー用のフレームワークを用いることでクローリングを行うことができる．

3.4　データベースツール

本研究ではマンガの推薦システムを行うためにスクレイピングとクローリングで抽出したデータからデータベースを作成する．MySQL は世界中で最も使われいているデータベースで高速で使いやすいことが特徴である．PostgreSQL は MySQL と同じぐらい使われているデータベースで, PostGIS という地図や幾何データを扱うことができる拡張がある．本研究では地図は使わないため, MySQL を用いてデータベースを作成する．

3.5　マンガデータベースサイト

漫画レビュー.com はマンガだけを扱うレビュー投稿サイトで, ランキングやマンガの評価, レビュー投稿などが行え, 登録マンガ数は 10,000 作品を超える．作品データベースはマンガやアニメなどに関する様々な情報, 評価, ランキング, レビュー投稿などの場を提供しているサイトで, マンガは 7,709 作品登録されている．メディア芸術データベースは文化庁が作成した, 2015 年までのマンガやアニメ, ゲーム, メディアアートが登録されたデータベースサイトである．マンガに関しては明治初期から 2015 年 12 月までに発刊されたマンガ単行本と, 同じく明治初期から 2015 年 12 月までに発行されたマンガ雑誌が登録されており, その数は 20,000 作品以上である．本研究では, ラフ集合を用いて推薦するシステムを構築するため, 作品登録数が多くレビューの記述が存在する漫画レビュー.com のサイトを利用する．

3.6 ラフ集合理論

ラフ集合 [4] は, 多くの対象の複数の条件属性と決定属性の値を示す決定表の解析に有用である. 対象を識別する条件となる対象の条件属性の集合と, 識別の目的となる決定属性の集合により表現した表を決定表と呼ぶ. また, 決定属性の属性値に基づいて分類される対象の集合を決定クラスと呼ぶ.

ラフ集合には, 条件属性により対象を同値類に分類の仕方に上近似と下近似の 2 つの方法がある. 上近似とは, ある属性の集合をとったときに, ある分類に帰属するか識別できない対象を含めた可能性集合を表す. 一方下近似は, ある分類に確実に帰属する対象を含めた確実性集合を表す. 本論文では, 好きと嫌いを確実に識別する下近似を表現する上で最小限必要な属性の集合を縮約と呼ぶ. 本研究では, 決定表から決定クラスに属する対象を確実に類別する縮約を抽出する. 縮約を用いると, 決定クラスを識別する最小限の属性とその属性値の組み合わせを抽出できる. その組み合わせは, 決定クラスを導き出す if-then ルール [5] として捉えることができる. 表 3 の決定表を用いて if-then ルールを抽出すると以下のルール (1)(2)(3) のようになる.

$$if\ B = yes\ then\ Y = \text{好き} \quad (1)$$
$$if\ A = yes\ and\ D = yes\ then\ Y = \text{好き} \quad (2)$$
$$if\ C = yes\ and\ D = yes\ then\ Y = \text{好き} \quad (3)$$

本研究では, 決定表から下近似による決定属性の if-then ルールの抽出を行う.

表 3 決定表の例

対象物	条件属性				決定属性
M	A	B	C	D	Y
M1	yes	yes	yes	yes	好き
M2	yes	no	no	yes	嫌い
M3	no	no	no	no	嫌い
M4	no	yes	no	yes	好き
M5	no	no	no	yes	嫌い
M6	yes	yes	yes	no	好き

4 実験

本章ではマンガリストの作成とラフ集合を用いた好みの推測手順について説明し, 実験結果と結果に対する考察を示す. 4.1 節では, クローリングと形態素解析を用いたマンガリストの作成について示す. 4.2 節ではラフ集合を用いた好みの推測手順について示す. 4.3 節では実験結果, 4.4 節では結果からの考察を示す.

4.1 クローリングと形態素解析を用いたマンガリストの作成

本節では, ラフ集合で計算を行うために必要なデータをクローリングと形態素解析を用いて収集する方法を示す. 今回は漫画レビュー.com というマンガデータベースサイトからレビュー抽出を行う. また, 漫画レビュー.com にあるカスタムランキングを用いてレビューが 20 件以上ある作品のみを表示するようにした.

ラフ集合で計算を行うために必要な特徴語と好みの度合いを決定する. レビューが 20 件以上ある 353 作品のレビューを 10 件ずつ集め, 1 つのテキストファイルにまとめた. また, 同時に各作品の総合点を抽出し, 1 つの CSV ファイルにまとめた. テキストファイルにまとめたレビューに Mecab を用いて形態素解析を行い, 名詞のみを抽出し出現頻度が多い順に並べた.

次に好みの度合いを決定するために先ほど収集した各作品の総合点から各作品の平均点を出し, 平均点以上を「好き」, 平均点未満を「嫌い」とした. それぞれの作品数は「好き」の場合が 218 件, 「嫌い」の場合が 135 件であった. 次に好みの度合いが 2 つの場合と 4 つの場合で正答率が変化するかを調べるために「好き」の中で再び平均点を出し, 平均点以上を「好き」, 平均点未満を「やや好き」とした. 「嫌い」も同じように平均点を出し平均点以上を「やや嫌い」, 平均点未満を「嫌い」と分別した. それぞれの作品数は「好き」の場合が 119 件, 「やや好き」の場合が 99 件, 「やや嫌い」の場合が 85 件, 「嫌い」の場合が 50 件であった.

4.2 ラフ集合による好みの推測

R 言語を用いるための RStudio 上でラフ集合を用いた好みの推測の正答率を出力する手順 [6] の一部を図 2 を用いて説明する.

(1) では R 言語に登録されている RoughSets のパッケージを使えるよう library 関数を用いてパッケージを読み込む.

(2) では sample 関数を用いて csv ファイルを読み込んだマンガリスト x をランダムに行列に保存し, 保存したものを data として保存する.

(3) では data を 8 割と 2 割のデータに分けるために round 関数を用いて丸みを行う.

(4) では SF.asDecisionTable 関数を用いて data の 8 割を一番右側の列が決定属性となるような決定表に変換し data.tra として保存.

(5) では data の残りの 2 割を用いて (4) と同様に決定表を作成し data.tst として保存. その際, 決定属性の列を削除して好みがわからないように作成する.

(6) では (5) で削除した決定属性だけの表を true.classes として保存する.

(7) では data.tra の決定表を用いて if-then ルールを抽出する.

(8) では抽出された if-then ルールを用いて data.tst のそれぞれのマンガの好みを予測する.

(9) では予測された好みが元の好みとあっているか比較し割合を示す.

```
> library(RoughSets) (1)
> data <- x[sample(nrow(x)),] (2)
> idx <- round(0.8*nrow(x)) (3)
> data.tra <- SF.asDecisionTable(data[1:idx,],
decision.attr=16,indx.nominal=16) (4)
> data.tst <- SF.asDecisionTable
(data[(idx+1):nrow(data),-ncol(data)]) (5)
> true.classes <-
data[(idx+1):nrow(data),ncol(data)] (6)
> rules <- RI.LEM2Rules.RST(data.tra) (7)
> pred.vals <- predict(rules, data.tst) (8)
> mean(pred.vals == true.classes) (9)
```

図 2 ラフ集合を用いた好みの推測手順の一部

4.3 実験結果

条件属性が 10, 15, 20, 25, 30 の決定表を用いて好みが 2 段階評価と 4 段階評価の好みの推測結果を示す. もとの決定表を 8 割と 2 割に分けるパターンを 10 通り用意しそれぞれ正答率を検出し平均値を求めた. 求めた平均値をまとめた表を表 4 に示す. 2 段階評価時の正答率の平均値のグラフを図 3, 4 段階評価時の正答率の平均値のグラフを図 4 に示す.

表 4 条件属性の数に対する正答率の平均値

	正答率	
条件属性	2 段階評価	4 段階評価
10	0.597183	0.283099
15	0.61831	0.280282
20	0.615493	0.277465
25	0.621127	0.290141
30	0.580282	0.316901

4.4 考察

2 段階評価時の正答率はおよそ 6 割程度, 4 段階評価時はおよそ 3 割程度の値を示した. 好みを 4 段階に分けた場合は信頼性に大きく欠ける. しかし表 4 と図 4 から条件属性が 25 と 30 の時 3% から 4% 増加したため, 条件属性が多くなると正答率も増加する可能性が得られた. 一方好みを 2 段階に分けた場合は 4 段階に分けた場合より 2 倍ほど正答率が上がっていることから信頼性の高さがうかがえる. しかし, 2 段階に分けた場合でも正答率が 6 割ほどしかないため実装するには厳しい部分がある. また, 2 段階評価では 4 段階評価のときとは違い条件属性が増えるにつれて正答率が下がっていることから推薦システムを実装する場合は条件要素を 15〜25 で設定することが勧められる.

5 むすび

本研究では, マンガのレビューを用いて決定表を作成しラフ集合により if-then ルールの抽出を行い, 抽出した if-then ルールを用いて好みの推測を行った. 今後の課題として, アンケートを通して好みの推測が正しく検出できているかの検証, ストーリー以外にも絵柄の好みの度合いを

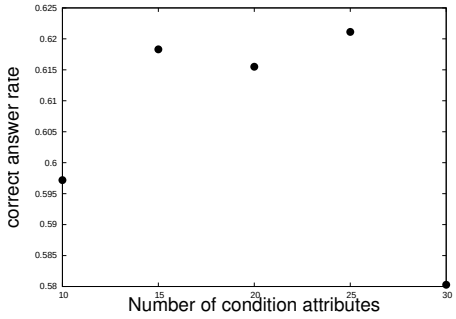

図 3 2 段階評価での要素それぞれの正答率の平均

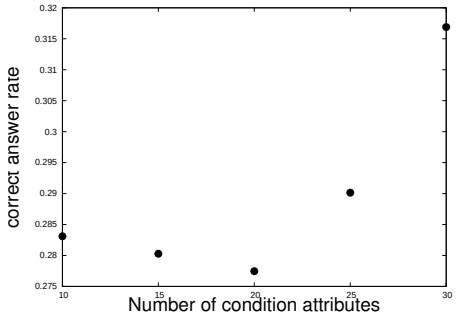

図 4 4 段階評価での要素それぞれの正答率の平均

考慮, 「世界」⇔「日本」, 「先生」⇔「生徒」といった選んだ特徴語と対極の特徴語の選出が挙げられる. また, 2 段階評価と 4 段階評価での正答率を向上するために条件属性の抽出時に tf-idf を行うことで条件属性それぞれに重みづけが可能となり, 正答率の向上につながると思われる.

参考文献

[1] 公益社団法人全国出版協会, “2017 年版出版指標年報,” 出版科学研究所, 2017.

[2] 山下諒, 朴炳宣, 松下光範, “コミックの内容情報に基づいた探索的な情報アクセスの支援,” 人工知能学会論文誌, Vol.32, No.1, pp.W2-D1-11, 2016.

[3] 西澤健吾, 萩野晃大, 中島伸介, “ラフ集合を用いた感性のモデル化に基づく推薦手法の提案,” (DEIM Forum2014), P2-1, 2014.

[4] Z. Pawlak, “Rough Sets,” International Journal of Information Computer Science, Vol.11, No.5, pp.341-356, 1982.

[5] 宮島卓也, 乾口雅弘, 鶴見昌代, 谷野哲三, “複数の決定表のラフ集合解析に関する基礎的考察,” 数理解析研究所講究録, Vol.1409, pp.234-247, 2005.

[6] Lala Septem Riza, Andrzej Janusz, Christoph Bergmeir, Chris Cornelis, Francisco Herrera, Dominik lezak, Jos Manuel Bentez, “Implementing algorithms of rough set theory and fuzzy rough set theory in the R package “RoughSets”,” Information Sciences, Vol.287, pp.68-89, 2014.