4B-6

閲覧ログのクラスタリングによる電子コミックのカテゴリ推定

佐藤 哲 †

NHN PlayArt 株式会社 ゲーム科学研究室 †

1 はじめに

弊社では電子コミックを提供するサービスを行っているが，電子コミックは男性/女性用，アクション/ギャグ漫画といった区別が難しいため，カテゴリ分類は敢えて行っていない．しかしユーザへのレコメンデーション，漫画を探す際の利便性を考えると，何かの規準に基づいたカテゴリごとに電子コミックが分けられているとユーザにとってもユーザデータの分析をする側にとっても有益である．そこで本研究では，ユーザの電子コミック閲覧ログをクラスタリングすることで，ユーザや電子コミックのデータマイニングを行う手法を提案する．

2 App/Web 用電子コミック

電子コミックと呼ばれる媒体は，既に非常に多くのサービスや製品があるため ††，詳細は論じない．ここでは，課金・非課金を問わず，スマートフォンアプリケーション (App) または Web ブラウザ (Web) を用いて閲覧する電子コミックを対象とする．その一般的な特徴として，書籍の漫画雑誌のように対象ユーザを限定していないことがあげられる．

弊社のサービス（図 1）でも，様々なコミックを一つの画面の中で提示している．男性用コミック，女性用コミック，アクションコミックなどの分類は提示していない．従って，画面を見ただけでは簡素な画像と短いキャプションが読めるのみで，好みの漫画を探すことが困難であることは明らかである．そこで自動分類・自動特徴抽出が必要となる．

3 ログデータ構造及び分析方法

ユーザの電子コミック閲覧ログは，通常の Apache httpd ログ形式である．すなわち，ユーザを匿名で識別するためのセッション cookie 情報やアクセス時間，アクセス先のアドレスなどが含まれている．電子コミックへのアクセス履歴は，次のようにベクトル化する．

$$v = (\delta_1, \delta_2, \cdots, \delta_n)$$

ただし，

$$\delta_i = \begin{cases} 1 & \cdots \quad \text{電子コミック } i \text{ を閲覧した} \\ 0 & \cdots \quad \text{電子コミック } i \text{ を閲覧しなかった} \end{cases}$$

Electronic Comics Category Estimation by Clustering User's Browsing Logs

†Tetsu R. Satoh, NHN PlayArt Corporation

††http://ja.wikipedia.org/wiki/Category:電子書籍

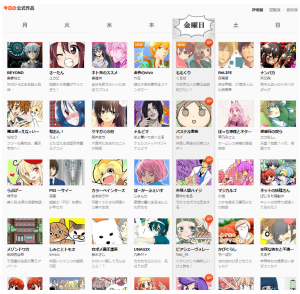

図 1: 電子コミックサービスサイトの例

であり，n は電子コミックの数である．例えば，$n = 2$ すなわち X 軸上に表されるコミック X と，Y 軸上に表されるコミック Y の 2 冊のみが対象とする．この場合，1 ユーザに対して起こりうるパターンは次の 4 種類である．左端から，何も閲覧されない状態，コミック Y のみ閲覧されていた状態，コミック X のみ閲覧されていた状態，コミック X とコミック X の両方を閲覧されていた状態を表す．要するに，$(0,0)$，$(1,0)$，$(0,1)$，$(1,1)$ の 4 種類のベクトルが構成される可能性がある．ここでもし，コミック X のみを閲

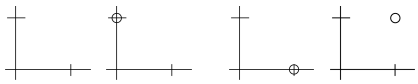

図 2: 閲覧データのベクトル化

覧したユーザと，コミック X 及びコミック Y の両方を閲覧した 2 名のユーザがいた場合，2 ユーザのクラスタリング結果は図 3 のようになるであろう．本

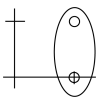

図 3: クラスタリング例

研究では，59 タイトルの作品を対象としたため，59 次元のベクトルをクラスタリングすることで知識発見を試みる．

クラスタリングは次のように行う．使用したシステムは，CDH 4.4，Hadoop 2.0.0，Mahout 0.7，Ruby

2.0.0，ログは2013/12/01から2013/12/31の一ヶ月分で，非圧縮の状態で容量は約8.3Gバイト，レコード数（行数）にして約3000万行である．計算環境は，ネームノード，ジャーナルノード，ヒストリーサーバ，データノードなど全てXeon L3426 8Core 16Gバイトメモリマシンで，データノードは3台である．まず，Rubyスクリプトによる Hadoop Streamingで，Webのログからユーザ IDと閲覧頁のペアの集計を行う．その結果，59次元空間のベクトル集合ができるので，それを Mahoutの Canopyアルゴリズム [1]を用いてクラスタリングを行う．最後に，その結果を研究者が分析する．

4　実験結果

Canopyアルゴリズムはクラスタの数を指定する必要が無いため，入力ベクトルの数と同じもしくはそれ以上の数のクラスタを構成することも可能である．ただし紙面の都合により，試した実験のうち一種類のパラメータ t1及び t2（Canopyアルゴリズム結果の，クラスタの大きさ及び数を決めるパラメータ）による結果のみを紹介する．

表1は，Canopyアルゴリズムを実行する際に与えたパラメータ及び実行結果である．ベクトル間の距離の計算には，デフォルトの2乗ユークリッド距離を用いた．ベクトルの要素数は59個のため，2つのベクトルの最大距離は，全要素がゼロのベクトルと全要素の値が1のベクトルの距離で59となる．そこでクラスタの半径を表すパラメータ t1は最大距離の1割前後を目安に複数の値で試行した．Mahoutによるクラスタリングの実行時間は約2時間であった．

表 1: 設定値/結果値

パラメータ	設定値/結果値
t1	5.0
t2	2.5
検出クラスタ数	36
クラスタ中最大ベクトル数	13695
クラスタ中最小ベクトル数	33

まず，最大要素数の $n = 13695$ であるクラスタ0について検討する．要素の重要度を表すクラスタ中心の座標値の大きさで，値が0.2を超えるものは「ReLIFE（図4左上）」† のみであった．クラスタ形状の歪さを表す特徴ベクトルの平均要素数は10タイトルコミックであり，要素が表すコミックの重要度も比較的高いもので，人気の高いものを広く読んでいる読者層のクラスタであることが分かる．このクラスタに属するコミックは「白ポメ高圧漫画 (図4右上)」††,「百獣のたてがみにゃんこ（図4左下)」†††,「ネト充のススメ (図4右下)」†††† など比較的ランキング上位にあり，広く人気を博するカテゴリに属するコミックであると言える．

次に，要素数 $n = 382$ のユーザが少ないクラスタについて考察する．重要度の高いコミックは「らぶ

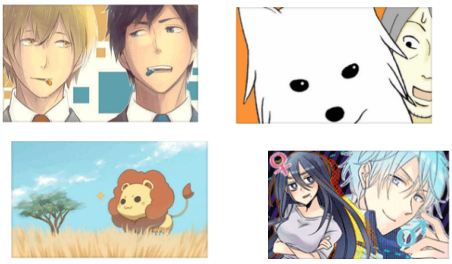

図 4: クラスタ0

図 5: クラスタ11

げー (図5左上)」†,「BEYOND(図5右上)」††,「『二重人格彼女』(図5左下)」†††,「魔法使＜えな＞いカエイクン (図5右下)」†††† などで，ランキングの上位下位には無関係に様々な作品が含まれていた．特徴ベクトルの平均要素数は11.6タイトルコミックで，人気があるコミックに限らず広く閲覧されているクラスタであると推定できる．ランキングが下位のコミックも含まれているという意味では，いわゆるマニアックなユーザに好まれるクラスタであると言える．

5　おわりに

ユーザの電子コミックに対する閲覧情報を分析することで，広く好まれるコミックなのか，少数のファンを持つコミックなのかなど多くの情報が得られることが分かった．特に，人気のあるコミックにユーザが集中するのは予想出来ていたが，比較的閲覧数の少ないコミックを好むユーザは，他にも閲覧数が少ないコミックをも好んで閲覧する傾向があるなど新たな発見もあった．今後は閲覧数だけではなく，画像やストーリーのタイプなどを自動的に考慮する高度なクラスタリングを行いたい．

参考文献

[1] A. McCallum, K. Nigam, and L. H. Ungar, Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching, Proc. 6th Int. Conf. Knowledge Discovery and Data Mining(SIGKDD), pp. 169–178, 2000.

†http://www.comico.jp/articleList.nhn?titleNo=2
††http://www.comico.jp/articleList.nhn?titleNo=20
†††http://www.comico.jp/articleList.nhn?titleNo=25
††††http://www.comico.jp/articleList.nhn?titleNo=27

†http://www.comico.jp/articleList.nhn?titleNo=38
††http://www.comico.jp/articleList.nhn?titleNo=59
†††http://www.comico.jp/articleList.nhn?titleNo=30
††††http://www.comico.jp/articleList.nhn?titleNo=55