F-016

スマートフォンマンガアプリのアクセスログ解析による ユーザ嗜好とコンテンツクラスタ関係の推定

佐藤 哲 †

NHN PlayArt 株式会社 データ研究室 †

1 はじめに

弊社では，スマートフォンアプリとして無料のマンガを提供するサービスを展開している．マンガのコンテンツはジャンルも対象ユーザも定めてはいなく，多様である．それゆえ，人気のあるコンテンツの判断，ユーザへの提供コンテンツのパーソナライズなどサービス品質向上のために必要な情報をログデータから判断することが難しい．

そこで本研究では，ユーザのアクセスログをクラスタリング/可視化することでサービスをユーザに提供する上で有益な知識発見をする手法を報告する．

2 アクセスログに対する閲覧ベクトル

弊社で提供しているスマートフォンマンガアプリは，サーバ側では一般的な Web サーバのアクセスログの形式でログが記録されている．アクセスログからは，ユーザを識別するための ID 及び閲覧しているマンガの ID が抽出でき，それを次のようにベクトル化する．ID=i のマンガと ID=j のユーザに対し，

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

ただし，

$$\delta_i = \begin{cases} 1 & \cdots \quad \text{マンガ } i \text{ を閲覧した} \\ 0 & \cdots \quad \text{マンガ } i \text{ を閲覧しなかった} \end{cases}$$

であり，n はマンガの数である．このベクトルはユーザの数だけ生成され，そのベクトル集合から特徴や傾向を調べることが本研究の目的である．便宜的に，このベクトルを閲覧ベクトルと呼ぶことにする．

3 処理概要

本研究では，2 段階の処理を施すことでアクセスログを解析する．

(1) 閲覧ベクトルをクラスタリングし，似た閲覧傾向のあるユーザをグルーピングする
(2) 全てのクラスタ間の類似度を計算し，無向グラフの形でクラスタ間のつながりを可視化する

前半の処理で，クラスタ内のユーザ同士で閲覧しているマンガを推薦しあうことでレコメンドサービスが実現できる．ただし，いわゆるパーソナライズし過ぎると意外性が無くなるという問題が発生するので，後半の処理で抽出した逆に類似度の低いクラスタの情報も参考にすることを考えた．使用したシステムは，ログデータが Hadoop クラスタの HDFS に保存されていることから Hadoop プラットフォームが中心であり，CDH 4.4，Hadoop 2.0.0，Mahout 0.7，Ruby 2.0.0 などである．ログは 2014/4/1 から 2014/4/7 の一週間分で，非圧縮の状態で容量は約 18.2G バイト，レコード数（行数）にして約 8000 万レコードである．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 24G 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 と閲覧したマンガ ID のペアの集計を行う．今回の実験では 59 タイトルのマンガを対象としたので，ユーザ ID+59 次元の閲覧ベクトルが出力される．そして生成された 59 次元空間のベクトル集合を Mahout の Canopy アルゴリズム [1] を用いてクラスタリングを行う．その後，クラスタ間の類似度をコサイン距離を用いて計算する．類似度計算にも Ruby スクリプトによる Hadoop Streaming を用いている．最後に，クラスタリング結果とクラスタ間の類似度を元に，Graphviz† を用いて可視化する．

4 実験結果

Canopy クラスタリングの設定値及び出力結果を表 1 に示す．距離の計算には二乗ユークリッド距離を用い，クラスタリング処理にかかった時間は約 100 分であった．

クラスタ間の類似度は，クラスタの中心座標のベクトル同士でコサイン距離を計算することで算出した．そして類似度が 0.3 より小さいペアは赤い実線で，0.6 より小さいペアは青い破線で，0.6 以上の類似度を持つペアは黒い点線でエッジを描画し，3 通りに大きく分類して可視化した．その全体像を図 1 に示す．また，クラスタ間類似度の異なる 3 種類のエッジが特徴的に現れている一部を図 2 に示す．図より，ID=4 のクラスタは，赤い実線のエッジのみで

表 1: 設定値/結果値

パラメータ	設定値/結果値
t1	6.0
t2	2.0
検出クラスタ数	31
クラスタ中最大ベクトル数	20699
クラスタ中最小ベクトル数	111

Estimation of User Preference and Contents Cluster Relation for Smartphone Comic Application
†Tetsu R. Satoh, NHN PlayArt Corporation

†http://www.graphviz.org/

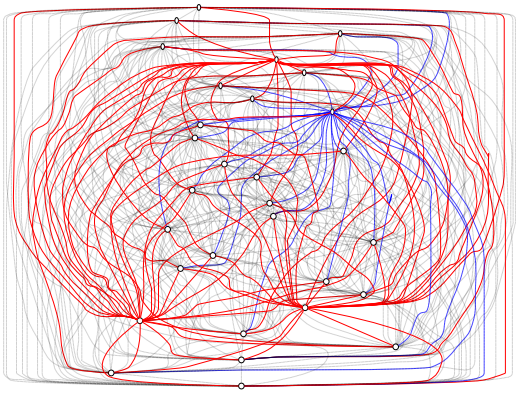

図 1: クラスタグラフ全体像

接続されており，他のクラスタとくらべて類似度が低い孤立したクラスタであることが分かる．ID=5 はほとんどが黒い点線のエッジであることから他のクラスタと傾向は似ており，ID=8 はその中間である．

Canopy クラスタリングでは，クラスタは中心と半径で表される．本研究では各座標が各マンガタイトルを表しているので，クラスタの中心の座標を調べることで，どのマンガタイトルに人気があるのか，人気が集中しているのか分散しているのかなどを調べることができる．そこで，ID=4,8,5 について，中心座標を確認したのが図 3 である．グラフは横軸が各マンガタイトルで，各タイトルにつき 3 クラスタの中心座標が縦軸を値としてプロットされている．従って，横軸の 1 メモリに対し，ID=4,8,5 の 3 つのクラスタの対応するベクトル成分の値がプロットされている．その結果分かることは，ID=4 の孤立クラスタは，複数のマンガについて明らかに縦軸の座標値で 0.1 以上，他のクラスタよりも高い値を示している．また，グラフからは分かりにくいが，縦軸の値が低いマンガタイトルについては，ID=4 のクラスタが ID=8,5 のクラスタよりも低い値を示している．また，ID=8,5 の両クラスタについては，厳密な統計的なチェックは行っていないが差はほとんど見られない．つまり，孤立クラスタは何かしら特殊な嗜好があるクラスタなのではないかと予想していたが

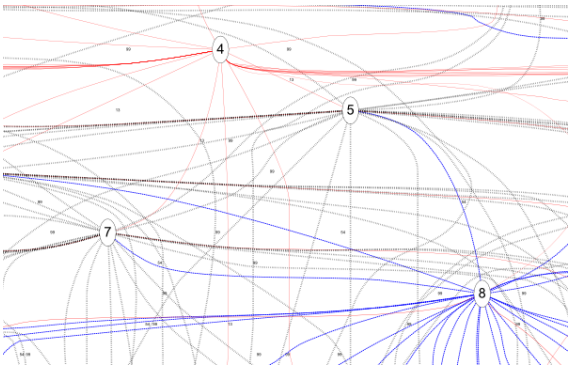

図 2: クラスタグラフ部分

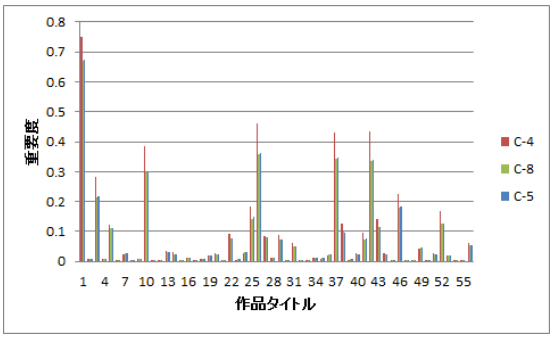

図 3: 漫画タイトル別重要度比較

（前回の報告 [2] では，孤立クラスタは一部の偏ったマンガを閲覧しているユーザだったり，マンガのイラストを特に重視するユーザである傾向があった），ID=4 の孤立クラスタのユーザは実際には読みたいマンガを読んで読みたくないマンガは読んでいないという自分の趣味に基づきマンガを読んでいるユーザであると思われる．言い換えると，ID=4 のクラスタのユーザが読んでいないマンガは，ユーザの好みでは無いか，コンテンツが魅力的ではないかのどちらかだと考えられる．そして結論としては，ID=4 のクラスタのユーザが読んでいないマンガは他のクラスタのユーザも読んでいないことから，ID=4 のクラスタのユーザは魅力的なマンガコンテンツに良い反応を示し，そうではないマンガコンテンツにはあまりアクセスしない，言わば「見る目のあるユーザ」ではないかとの仮説が立てられる．ユーザ単位での分析ができていないので，この仮説はいまのところ実証できていなく，今後の課題である．また，この簡単な考察では類似度が低いクラスタ同士でも全体の傾向は似ており，そのまま意外性の発見に利用することは難しいことが分かった．

5 おわりに

本研究は，情処全大にて発表した研究 [2] の第二報といえる．そのため研究結果の比較を試みることを考えていたが，ユーザの大幅な増加のため，ログ量の増大による分析負荷，ユーザ数増大によるクラスタ間の差の減少など多くの変化があり比較研究はできなかった．しかし，ユーザ数が増大するとアクセスマンガタイトルが似通ってきてクラスタ分析が難しくなることが分かるなど，多くの知見が得られた．本稿の図は理解し難いと思われるため，発表当日に詳しく説明する予定である．

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.

[2] 佐藤哲, 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, 第 76 回情処全大, 4B-6, 2014.