D-025

Hadoop を用いた電子書籍ユーザのアクセスデータ分析

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を使ってマンガやノベルを無料で読むことができるサービスを提供している．配信コンテンツは 4500 作品を越えており ††，その量の多さから，様々な切り口でのレコメンドやクラスタリングが必要となっている．そこで本発表では，Hadoop を用いてユーザのアクセスログを分析する手法の 1 つを提案する．発表の中では，コンテンツ配信アプリケーションのログデータに適するユーザアクセス DNA シーケンスを定義し，それによりユーザ間の類似度を計算してユーザをクラスタリングをした結果を紹介する．

2 ユーザアクセス DNA シーケンスと類似度計算

サービスのアクセスログには，リクエスト毎にアクセス時間やアクセス先 URI などの情報が記録されている．その中から，ユーザアクセス DNA シーケンスの定義に必要な以下の情報を抽出する:

(1) アクセス時間
(2) ユーザ識別 ID
(3) 作品を識別するマンガタイトル ID
(4) 第何話かを示すマンガ話数

これらをテキストとして羅列したものを考える．例えば “150601101430T9333A4” は，2015 年 6 月 1 日 10 時 14 分 30 秒にタイトル ID9333 の作品の第 4 話を閲覧したことを意味する．そしてこの記号列をカンマ区切りでユーザ毎に時系列に羅列したものをユーザアクセス DNA シーケンスと定義し，その例を図 1 に示す．また，お知らせやトップページなどの配信コンテンツ以外へのアクセスではマンガタイトル ID やマンガ話数が発生しないので，ユーザアクセス DNA シーケンスの中にマンガタイトル ID とマンガ話数が存在しないユーザはコンテンツにアクセスしていなく，不正アクセスまたはシステム不具合の可能性がある．このような場合のユーザアクセス DNA シーケンスを図 2 に示す．

ここで定義したユーザアクセス DNA シーケンスに対し，NCD(Normalized Compression Distance)[1]

図 1: DNA 例

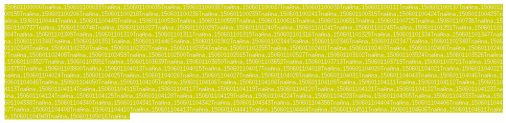

図 2: 特異ユーザの DNA 例

を適用し，ユーザ間の類似度を計算する:

$$NCD(x,y) = \frac{Z(xy) - \min(Z(x), Z(y))}{\max(Z(x), Z(y))}$$

ここで，$Z(x)$ は文字列 x を圧縮した後の長さであり，$Z(xy)$ は文字列 x と y を結合して圧縮した後の長さである．NCD は距離であるので，NCD の値が小さいほど類似度が高くなる．

類似度計算を MapReduce を用いて実現するために，次のようなステップで処理を実行する．

(1) ログをパースし，ユーザ毎のアクセスログを生成する
(2) ユーザ毎のアクセスログから，ユーザ毎のログの圧縮サイズを計算する
(3) ユーザ毎の圧縮後ログサイズより NCD を用いてユーザ同士の類似度を計算する

処理 (1) のみ，パース結果をユーザ毎に集約する必要があるので，map と reduce を使用している．他の処理は map のみで実現可能である．

3 実験結果

前節で述べた類似度計算結果を元にアクセスデータが類似していると判断できるユーザをクラスタリングすることでユーザの分析を試みる．図 3 に，階層的クラスタリングによりユーザをクラスタリングした結果の例を示す．使用したのは 2015 年 6 月 1 日の午前 10 時台の 1 時間のログで，クラスタリングには最低限の量のログのが記録されている 103 ユーザ

E-Book Access Data Analysis Based on Hadoop
†Tetsu R. Satoh, NHN PlayArt Corporation
††http://www.nhn-playart.com/press/index.nhn?m=read&docid=8398952

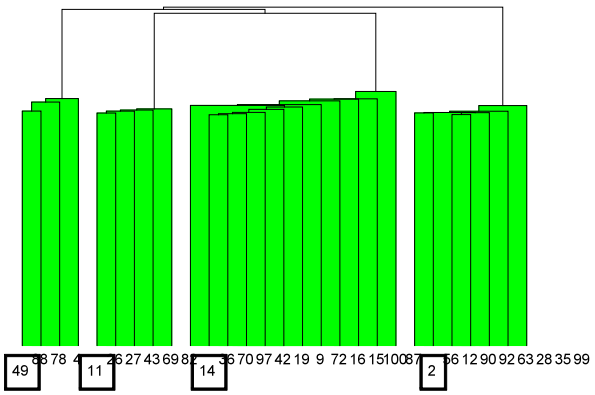

図 3: 階層的クラスタリング適用例

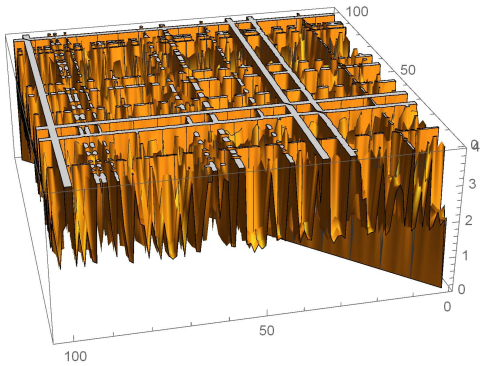

図 4: ユーザ間の距離の計算例

を対象とした．横軸の数値はユーザを識別する ID 番号であり，縦軸の線の長さはクラスタ同士の距離に対応している．この図から，大きく 4 つのクラスタに分かれていることが見て取れる．この場合のユーザ間の距離行列の値を図 4 に示す．縦横軸はユーザを識別する ID 番号で，高さが NCD の値である．本来，正規化されている距離に大きな値が出ていることから，データが少ない等の理由で NCD がうまく計算されていない場合があることが分かる．この問題は，ユーザアクセス DNA シーケンスの定義にも由来するので，符号化方法は今後改良していく必要がある．データ量の少なさは，使用するログの量を増やすことである程度解決ができる見込みである．

次に，各クラスのいくつかの最下位ノードを取得し計算した，クラスタの特徴を表すと考えられる特徴量を表 1 に示す．ここで，平均閲覧ページとは取得したログの中でユーザがコンテンツにアクセスした回数をユーザ数で割った値であり，重複閲覧タイトル割合は複数のユーザが同じ作品を閲覧した割合，連続閲覧タイトル割合はユーザがある作品を 3 話以上読み進んでいた割合である．C1 から C4 までの記号はそれぞれクラスタ 1 からクラスタ 4 を表している．つまり，平均閲覧ページはクラスタに属するユーザがヘビーユーザなのかライトユーザなのかを表し，重複閲覧タイトル割合はクラスタ内のユーザの嗜好の類似度を表す．連続閲覧タイトル割合は，ユーザが気に入った作品があったかどうかなどの情報が含まれると仮定している．連続閲覧タイトル割合が少なく平均閲覧ページが多いクラスタは，定期的に更新作品を閲覧する定常ユーザのクラスタである可能性があり，連続閲覧タイトル割合が多いクラスタのユーザは気に入った作品があるか，または気に入る作品を探している，新たに本サービスの利用を始めた，等の理由から作品をまとめて読んでいるユーザである．

表 1: 各クラスタ最下位ノードサンプルの特徴量

パラメータ	C1	C2	C3	C4
平均閲覧ページ	192	103	94	49
重複閲覧タイトル割合	21%	11%	21%	7%
連続閲覧タイトル割合	8%	17%	17%	19%

クラスタ 1 は，平均閲覧ページも重複閲覧タイトル割合も多く，嗜好が類似するヘビーユーザが属するクラスタになっている．連続閲覧タイトル割合が低いので，定常的なユーザであることも予想される．クラスタ 2 とクラスタ 3 を比較すると，重複閲覧タイトル割合の値からクラスタ 2 の方が好みが分かれている可能性がある．クラスタ 4 は平均閲覧ページが少ないことから，ライトユーザのクラスタの可能性が高い．ただしクラスタ 4 に分類されたユーザの中にも連続閲覧タイトルが多いユーザもいるため必ずしもライトユーザだけのクラスタとは限らない．類似度計算が不正確なのか，他のクラスタとは違う特性のユーザが分類されているのかなどの面で検討している．

以上の実験では，MapReduce の実行には Ruby スクリプトによる Hadoop Streaming を，計算結果の格納には HBase を，NCD の計算に用いる圧縮アルゴリズムは bzip2 を，階層的クラスタリングには Wolfram *Mathematica* を用いた．クラスタリングは，クラスタに属するノード同士の距離の最小値を用いてクラスタ結合が行われている．

4 おわりに

本発表では，電子書籍ユーザのアクセスログに対しユーザアクセス DNA シーケンスを定義し，それに対し NCD を適用して類似度を計算することを提案した．そして類似度に基づき階層的クラスタリングを用いてユーザをクラスタリングした結果を紹介した．

参考文献

[1] P. M. B. Vitányi, Compression-Based Similarity, Proc. Int. Conf. Data Compression, Communications and Processing, pp. 111-118, 2011.