F-020

アクセスパターンのクラスタリングによる Web ログからのユーザ属性推定

佐藤 哲 †
NHN テコラス株式会社 データ研究室 †

1 はじめに

インターネットを利用したサービスが日常的に利用されている現在，Web のログからのデータマイニングは多くの研究テーマが存在し注目を集めている．特に企業においては，Web を介したサービスを利用しているユーザのサービスに対する好みや満足度を推定したり，また満足度の低いユーザはどのようなユーザなのか，ユーザの好み，年齢，性別など様々な属性を推測することはサービス改善のために意味のある課題である．そこで本研究では，教師無し学習であるクラスタリングを用いてユーザ属性を推定する手法について述べ，予備実験結果を紹介する．

2 ログの ETL 処理と類似度計算

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元 IP アドレス，ブラウザ名やバージョン，アクセス先 URI などの情報が記録されている．その中から，ユーザのアクセスパターンを把握するために必要な，(1) アクセス時間，(2) ユーザ識別 ID，(3) アクセス先ページ ID，の情報を抽出する．これらをテキストとして羅列したものを考える．ここで，アクセス時間とアクセス先ページ ID は共に数字で表され，後述する類似度計算において不都合を生じるため，アクセス先ページ ID は文字コードを変換してアルファベットに置き換える．例えば “16021110-HEDI-EG” は，2016 年 2 月 11 日 10 時台に，ID が HEDI-EG のコンテンツにユーザがアクセスしたことを意味する．そしてこの記号列をユーザ毎に時系列に羅列したものを作成し，類似度計算に用いる．本研究では，この Web ログから情報を抽出した文字列をユーザアクセス DNA シーケンスと呼ぶ．構成内容から，ユーザが Web コンテンツにアクセスする時間情報とコンテンツ情報が含まれるため，ユーザの行動時間及び嗜好情報に基づいたデータ分析が可能であることが期待される．

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] 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．そして全てのユーザ間の距離を計算し，距離に応じたクラスタリングを実施する．

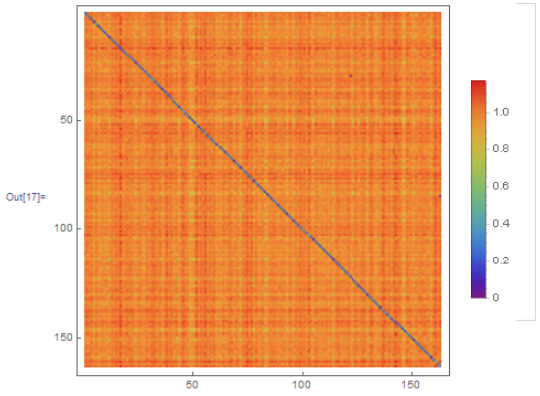

図 1: ユーザ間の距離の計算例

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．実験は Hadoop クラスタ上の Spark を用いており，ユーザ間類似度計算結果などの中間データは HBase に格納している．データは弊社のサービスのログで，一般的な Apache HTTP サーバ形式のログデータである．

まず，ユーザ間の距離行列の値の可視化結果を図 1 に示す．NCD の計算に用いる圧縮アルゴリズムは，bzip2 を用いた．縦軸，横軸がユーザ ID を表しており，色が青に近いほど距離が近く，赤に近いほど距離が遠いことを表している．対角線上は同じユーザ同士の距離に対応するので距離はゼロに近くなり，やや似ていることを表す黄色いエリアが現れていることが分かる．ただし，距離が 0.5 以下のエリアは確認できないが，その原因は判明していない．クラスタリングは，この距離行列を Spark MLlib†† に実装されている PIC(Power Iteration Clustering)[2] を用いて実施する．

図 2 に，PIC 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 2016 年 2 月 11 日（木）から 2016 年 2 月 13 日（土）のログで，クラスタリ

User Properties Estimation from Web Logs by Clustering Access Patterns
†Tetsu R. Satoh, NHN Techorus corp.

††http://spark.apache.org/mllib/

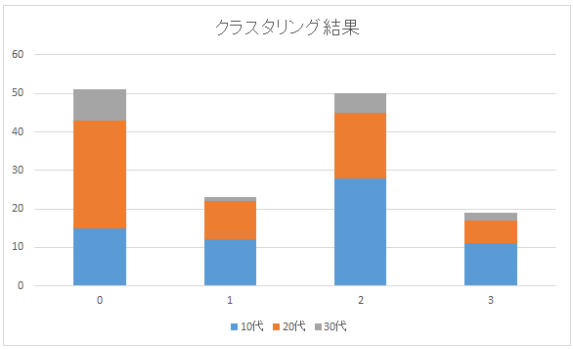

図 2: PIC 適用例

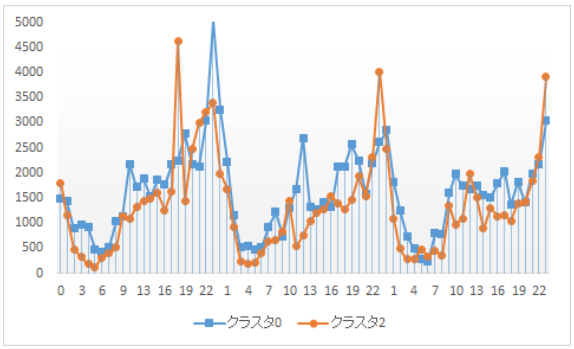

図 3: クラスタ 0 とクラスタ 2 の時系列アクセス数

ングには十分の量のログが記録されており，かつ属性がある程度わかっている 157 ユーザを対象とした．期間は，データ量，以前の研究との比較及び平日と週末の変化を調査する目的により決定した．横軸はクラスタ ID 番号であり，縦軸はそのクラスタに属するユーザ数である．また，PIC は k-means クラスタリングも内包しており，従ってクラスタ数は事前にパラメータとして与える必要がある．本稿で紹介する結果は，クラスタ数 k=1,2,3,4 を試したのち，妥当性と過去の研究結果の比較を目的としてクラスタ数 k=4 の結果を紹介している．特に情報量規準や尤度などを利用して決定したわけではない．今回対象とした 157 ユーザは年齢情報がある程度分かっているため，10 代，20 代及び 30 代のユーザをカウントし，色分け表示してある．

クラスタリング結果を考察すると，クラスタ 0 及びクラスタ 2 は所属ユーザ数が約 50 と酷似しており，クラスタ 1 及びクラスタ 3 がやはり所属ユーザ数が約 20 と似ている．従って，クラスタ 1 とクラスタ 3 はほぼ同様の属性を持つユーザのクラスタと考えマージすることが可能であろう．

ここで興味深いクラスタ属性は，クラスタ 0 は 20 代のユーザが多く，クラスタ 2 は 10 代のユーザが多い点である．この点を分析するため，図 3 に示すように，二つのクラスタの時系列的なアクセス数を確認した．縦軸はアクセス数に対応し，横軸は 2016 年 2 月 11 日 0 時からの 72 時間の 1 時間毎の時間を表す．四角マーカ付き青色線がクラスタ 0 を，丸マーカ付きオレンジ色線がクラスタ 1 を表している．その結果，明白な違いがいくつか観察できる．まずクラスタ 0 は，深夜ピークの 3000 アクセス以上の約半分にあたる 1500 アクセスに達する時間が早い．クラスタ 1 が午前中から徐々にアクス数が増えて夕方から夜にかけて 1500 アクセスに達するのに対し，クラスタ 0 は午前中のうちから 1500 アクセスに達し，概ね夕方・夜までアクセス数が維持されている．これは，クラスタ 0 のユーザの方がクラスタ 1 のユーザに比べ，朝から自由にスマートフォン等を利用できるユーザであると推測できる．また，深夜・明け方帯アクセス数が，クラスタ 2 の方が明らかに減少している．これは，クラスタ 0 のユーザに比べクラスタ 2 のユーザの方が，深夜帯に活動しにくいことを意味すると思われる．その他，クラスタ 0 のユーザの方が，クラスタ 2 のユーザに比べ，週末のアクセス数が減少していることが見て取れる．これは，クラスタ 0 のユーザは週末には Web アクセス以外の行動が増加する傾向があることを示唆していると思われる．

以上のような 2 つのクラスタの特徴の差異は，20 代ユーザの方が午前中から自由に Web にアクセスできる，深夜も活動する可能性がある，週末に他の活動の関係で Web アクセスが減少する可能性がある，など一般的な感覚に即しているように思われる．これらの結果から，例えユーザの属性が不明だとしても，クラスタリングによりユーザの属性の一つである年齢層を推定できる可能性がある．

以上の実験では，データノード 6 台の Hadoop クラスタ上で，YARN リソース管理を用いた Spark 環境を利用した．ログサイズは 3 日間分で約 25 ギガバイトで，類似度計算及びクラスタリングの処理時間は 20 分以内で終了した．データノードのサーバスペックは，CPU Xeon E5-2630(2.00GHz)，メモリ容量は各 64G バイトである．

4 おわりに

本発表では，Web サービス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し，PIC によりユーザをクラスタリングすることで，クラスタに属するユーザの属性を推定する手法について提案した．そして類似度計算結果と PIC 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.

[2] F. Lin, W. W. Cohen, Power Iteration Clustering, Proc. 27th Int. Conf. Machine Learning, pp. 655–662, 2010.