ISS-P-136

大規模 Web セッション分析に向けた分割クラスタリングの性能評価

南　哲志† 　橘　完太††
†工学院大学情報学部コンピュータ科学科　　††工学院大学情報学部情報デザイン学科

1. はじめに

Web サイトに蓄積されたデータ解析の手法の一つとして，アクセスログや購入履歴など「利用した情報」に着目する Web Usage Mining [1]が盛んに研究されている．Web サイトを訪れたあるユーザーが，一定の間を空けず閲覧したページの集合はセッションと呼ばれている． Web Usage Mining ではセッション間の類似度を用いたクラスタリング [2] [3]などが行われている．しかし，先行研究では比較的小規模なサイトやアクセスログデータに対してのみしかクラスタリングを施せていない現状がある．原因の一つとして，類似度行列の計算に時間が掛かる事が挙げられる．類似度行列の計算量は，セッション数 s に対して$O(s^2)$であり，1 回の類似度計算が数ミリ秒で終わるとしても多大な処理時間が掛かってしまう．そこで本研究では，処理時間を減らすため，分割クラスタリングの手法を提案する．提案手法では，理論上，分割数 m ，クラスター数 c に対して計算量は $O(\frac{s^2}{m}+m^2c^2)$ である．更に後述の最適な分割数を選択すれば，計算量は$O\left(s^{\frac{4}{3}}\right)$ となる．最後に提案手法を一致度と計算速度の観点から評価する．

2. 提案手法と実験

図 1 に手法と実験の概要を示す．まず，対象となるアクセスログデータから， n アクセス程度のセッション集合 X を抽出する．ここで従来手法としてセッション集合 X に階層型クラスタリングを施す．次に提案手法として，セッションの集合 X から分割したデータ $Y_1 \ldots Y_m$ として n アクセスのデータを m 個に分割したデータを取り出す．続いて，セッション集合 $Y_1 \ldots Y_m$ にそれぞれ階層型クラスタリングを施す．その後，$Y_1 \ldots Y_m$ のそれぞれが持つ c 個のクラスター内で，代表セッションを導出する．代表セッション y_i^k $(1 \le i \le m, 1 \le k \le c)$ は，Y_i^k に属する他セッションとの類似度の総和が最も高くなるセッションであり，

$$y_i^k = \underset{y' \in Y_i^k}{\operatorname{argmax}} \left(\sum_{\substack{y \in Y_i^k \\ y \neq y'}} S_{sim}(y, y') \right) \quad (\mathrm{i})$$

で表される．代表セッション y_i^k の集合を Y' とする．ここで，セッション集合 Y' に関してもクラスタリングを施す．また，この Y' に関しても代表セッション $y^{\ell'}$ $(1 \le \ell' \le c)$ を求める．次に，セッション集合 Y' のクラスターとセッション集合 X のクラスターとの Confusion Matrix である行列 M を作成する．次に M に関して対角要素を最大化するように行と列を並べ替え，そのトレースを一致数とする．最後に一致数をセッション集合 X の長さで除した値を一致度とする．

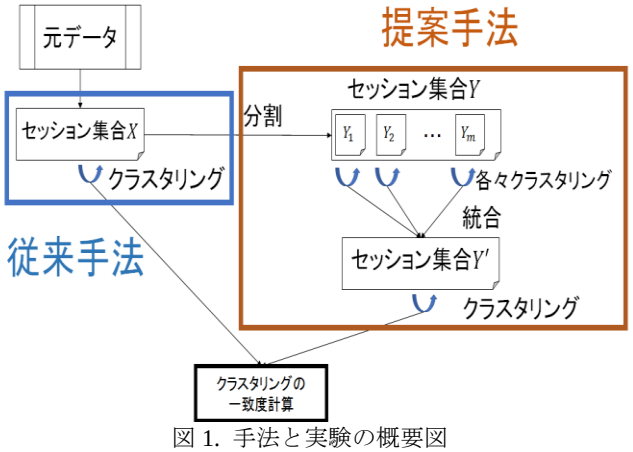

図 1. 手法と実験の概要図

3. 最適な分割数

ここで最適な分割数 m を求める．計算オーダーは $O\left(\frac{s^2}{m}+m^2c^2\right)$ であった．$m = s^p$ として考えると，計算オーダーは $O(\frac{s^2}{s^p}+s^{2p}c^2)$ となり，p が増加すると第二項の，p が減少すると第一項が計算オーダーとなる．ゆえに両項が釣り合う p を求める事で最良のオーダーが導出可能である．実際に p を求めると $p=\frac{2}{3}$ となり，この事から m は $s^{\frac{2}{3}}$ に比例して決定すればよい事がわかる．また最適な m を用いた計算量は $O(s^{\frac{4}{3}})$ となる事が明らかとなった．

4. 結果

今回の実験では，セッション集合 X の全 1,023 セッションの内，一致数は 907 セッション，一致度は約 0.89 となった．またクラスタリングの処理時間は約 85%の削減となり，クラスターを分割しても高速に，似ているクラスタリング結果を得られる事が明らかとなった．

参考文献

[1] M. Aldekhail. Application and Significance of Web Usage Mining in the 21st Century: A Literature Review.: International Journal of Computer Theory and Engineering, Vol. 8, No. 1, February 2016, p41-p47, 2016.

[2] Mrs. G. Sudhamathy, Dr. C. Jothi Venkateswaran . Web Log Clustering Approaches - A Survey.: International Journal on Computer Science and Engineering (IJCSE) Vol. 3 No. 7 July 2011 p2896-p2903, 2011.

[3] G. Poornalatha, Prakash S. Raghavendra. Web User Session Clustering Using Modified K-Means Algorithm.: ACC 2011, Part II, CCIS 191, pp. 243-252, 2011., 2011.