Webアクセスログのクラスタリングとその利用*

6X－02

大浦　勇亮†　喜連川　優‡
東京大学生産技術研究所§

1 はじめに

Webサイト構築者にとって、サイトを構築する際に、どの情報(Webページ等)をどこに配置すべきか、どのような階層構造(Webページのリンク階層等)にするのがユーザにとって利用しやすいのかを考えてサイトを構築することは非常に重要なことである。しかしながら、これは容易な作業ではない。そこで、サイト構造の問題点を発見し、それをサイト構造の再構築やユーザのナビゲーション支援へと利用するための、Webアクセスログによるユーザの行動パターンの分析に関する研究が盛んになってきた。本発表では、アクセスログを利用したユーザセッションのクラスタリング実験結果と、それを用いたユーザの問合せ拡張支援に関して検討を行う。

2 Webアクセスログのクラスタリング

ユーザセッション

アクセスログには、リクエストされたURLアドレス、アクセス日時、リモートホスト名をはじめとして、Cookieやフォームに入力された内容等を残すことが出来る。ログには不特定多数の人によるリクエストが全て記述されるが、Cookie IDが取得できている場合は、それによって同一ユーザのリクエストをログから特定することが出来る。同一ユーザでも訪れる度ごとにその要求は変化するため、通常、アクセスログの分析では、セッションという概念を導入する。セッションは、同一ユーザのリクエストにおいて、各リクエスト間隔が30分以内であるリクエストの集合によって表される。

アクセスログを用いて、どのWebページを閲覧したかに関してユーザセッションをクラスタリングする研究はすでにいくつか発表されているが[2][3]、本実験では、業種や店舗名を入力することによって店舗情報を検索することが出来る職業別電話番号情報サイト[1]のWebアクセスログを用いて、その問合せに入力された業種に関して、類似するユーザセッションのクラスタリングを行った。

クラスタリングアルゴリズム

まず、ユーザセッションをその入力した業種によってベクトル $\vec{s}$ で表す。ここで、業種総数を N_g とすると、i番目のセッション S_i のセッションベクトル $\vec{s_i}$ は、そのセッションにおいて業種jが入力された場合に $1(s_{ij}=1)$、入力されていない場合は $0(s_{ij}=0)$ を値にとる N_g 次元のベクトルとした。

クラスタリングアルゴリズムは、一般的に広く使われているK-meansアルゴリズムであるが、生成すべきクラスタ数が未知であるため、初期値としてクラスタ数Kを与えるのではなく、類似度閾値 TH_{sim} を与えてクラスタリングが行えるように改良した。また、類似度の計算にはコサイン値を使用した。このアルゴリズムでは、最初に選ばれる代表点間の距離が少なくとも類似度閾値 TH_{sim} よりも離れていることが保障され、異常値データは孤立クラスタとすることができるのが利点である。アルゴリズムの詳細は以下の通りである。

入力セッション数がNである時、入力ベクトル $\vec{s_1},\vec{s_2},\vec{s_3},,,\vec{s_N}$ に対して、

1. 最初の入力ベクトル $\vec{s_1}$ をクラスタ C_1 の中心ベクトル $\vec{c_1}$ とし、S_1 を C_1 のメンバとする。以後、$S_2,S_3,,,S_N$ では、入力を $\vec{s_i}$ とした時、既存のクラスタ $C_1,,,C_k$ との類似度を全て計算し、どのクラスタとの類似度も閾値 TH_{sim} 未満ならば、新たなクラスタを生成してそのクラスタ中心とし、類似度が閾値 TH_{sim} 以上ならば、最も類似度が高いクラスタのメンバとする。最後の入力 $\vec{s_N}$ まで割り当て終わったら、生成された全クラスタに関してその重心ベクトルを計算する。

2. 再び、入力 $\vec{s_1},\vec{s_2},\vec{s_3},,,\vec{s_N}$ に対して、既存クラスタとの類似度を計算し、閾値 TH_{sim} 未満の場合は、新たなクラスタを生成してそのクラスタ中心とし、類似度が閾値 TH_{sim} 以上の場合は、最も類似度が高いクラスタのメンバとする。

3. 最後の入力 $\vec{s_N}$ まで一通り割り当て終わったら、生成された全クラスタに関してその中心ベクトルを再計算する。

4. ステップ2,3を繰り返して、割り当てが変化しなくなったら終了する。

---

*Clustering of Web access logs and the application.
†Yusuke Ohura
‡Masaru Kitsuregawa
§Institute of Industrial Science, University of Tokyo

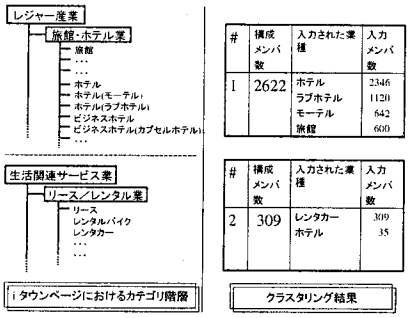

図 1: iタウンページにおける業種階層 (左) とクラスタリング結果 (右)

クラスタリング結果

今回、クラスタリングに使用したデータセットは1ヶ月分 (2000 年 5 月) の Web アクセスログであり、対象としたのは、Cookie ID がとれたユーザ (426,031 人) の検索リクエスト (5,094,913 件) のうち、2 つ以上の業種が入力されたユーザセッションを対象とした。この時、該当するセッション数は 125,847 であった。得られた結果を解釈するにあたって、最小クラスタサイズ MIN_{cl} よりも少ないメンバ数のクラスタはクラスタとしてカウントしないことにする。今回の実験では、この値を対象セッション数の 0.01% (MIN_{cl} = 12.58 人) 未満とし、それに相当する 12 人以下のユーザセッションで構成されるクラスタは切り捨てた。また、閾値 TH_{sim} には 0.10 を用いることにした。その結果、得られたクラスタ数は 667 個であった。

結果として、例えば" ホテル"," ラブホテル"," モーテル"," 旅館" を検索したクラスタや、" レンタカー"," ホテル" を検索したクラスタ等が得られた。この時、前者のクラスタは i タウンページにおける業種階層において同一階層にある業種から構成されるが、後者のクラスタには業種階層の異なる業種が含まれている (図 1)。この他にも同一階層にはない業種で構成されるクラスタが多数得られ、問合せを行う際に、サイト構築者の意図とは異なる意図を持つユーザが存在する可能性があることがわかった。

3 問合せ拡張支援システム実現の検討

クラスタリング結果を利用して、ユーザがサイトを利用するのを支援するアプリケーションの実現に関して検討する。動的にリンクを生成するシステム [2] 等はすでに発表されているが、我々は、ユーザの問合せに関して、入力すべき業種名を拡張支援するシステムの構築を考案する。

我々が考案する問合せ拡張方法は、以下の 2 ステップから構成される。

1. ユーザによって入力された業種が属するクラスタのうち、クラスタ構成メンバ数に対して、その業種を入力したメンバ数の割合が最も高いクラスタを利用して、業種の拡張提示を行う。
2. 入力された業種が属するクラスタのうち、第 1 ステップで使用していないクラスタにおいてトップにある業種と、サイトに導入されている業種階層とを比較して、入力された業種と同一階層になければその業種を提示する。

第 1 ステップでは、アクセスログに基づいて、業種を変更する際に使用する可能性の高い業種を提示する。例えば、" ホテル" が入力された場合、" ホテル" の割合が最も高いクラスタを用いて、" ラブホテル"、" モーテル"、" 旅館" を提示する。

第 2 ステップでは、さらに業種が異なる階層にあるものの、ログから" ホテル" と関連度が高いと判定された業種を提示する。これによって、" ホテル" の入力に対して、さらに" レンタカー" や" 結婚式場"," 貸会議室" 等を提示することが出来る。

4 おわりに

本稿では、i タウンページの Web アクセスログを用いて、ユーザセッションの問合せに関してクラスタリングを行った。その結果、用いられている業種階層とは異なる階層の業種を含むクラスタが得られ、ユーザの行動や意図を理解するのに役立つことがわかった。また、クラスタリングの結果を利用した問合せ拡張システムを考案し、ユーザの利便性の向上を図れることを述べた。今後は、この問合せ拡張システムの開発を進めるとともに、相関マイニング等の他の手法を用いた結果とクラスタリング結果とを比較検討する予定である。

謝辞

アクセスログを提供していただいているＮＴＴ番号情報株式会社ならびにご協力くださっている方々全てに深く感謝いたします。

参考文献

[1] i タウンページ http://itp.ne.jp/

[2] T. W. Yan, etc., From user access patterns to dynamic hypertext linking, *In Proc. of the 5th International World Wide Web Conference*, vol.28, 1007-1014, 1996.

[3] O. Nasraoui, etc., Mining Web Access Logs Using Relational Competitive Fuzzy Clustering, *In Proc. of the 8th International Fuzzy Systems Association World Congress - IFSA 99*, 1999.