

位置指向の Web ログマイニング

高橋 克巳^{†‡} Iko Pramudiono[†] 喜連川 優[‡]

[†] 日本電信電話株式会社 NTT 情報流通プラットフォーム研究所

[‡] 東京大学生産技術研究所

あらまし

Web のサービスへのアクセス記録を解析する Web ログマイニングにおいて、ログデータを地理的な位置にもとづいて解析する方法に関して述べる。Web で利用者の現在位置や利用者の望む位置に関する情報の提供サービスがさかんになってきており、効果的な解析方法と、有効な利用方法の考察が必要である。本稿では、インターネットでの位置に関する情報提供サービスの大規模ログデータを解析した作業の報告を行う。解析から、住所ワードの推薦には、commodity である情報対象に対する近隣地域の提示、および specialty に対する中心地域の推薦が可能であること、およびそれぞれのルールを選ぶ際には相関ルールの支持度だけでなく、単語間の相互情報量が有効であることについて述べる。最後に本解析結果を使った単語を推薦して検索の補助を行うシステムについて説明する。

Location Based Web Log Mining

Katsumi Takahashi^{†‡} Iko Pramudiono[†] and Masaru Kitsuregawa[‡]

[†] NTT Information Platform Laboratories, NTT Corporation

[‡] Institute of Industrial Science, The University of Tokyo

Abstract

A method of location based Web log mining is discussed. Recently web information provision based on the geographic location became more popular, location-based mining method and utilization of geographical rules is required. We describe the study of analyzing actual large Web log data in location based way. In this analysis, we found that we can distinguish geo-words rules in two purposes, nearby location provision for commodity objects and heartlands for specialty objects using support and mutual information of items. A recommender prototype of search words presentation to support users' searches is described as well.

1. はじめに

インターネットで入手可能な情報の量の増大に応じて、手にする情報の質の向上が求められるっており、このニーズを背景にして現在 Web マイニングの研究が行われているといえる。Web マイニングの中で、Web のサービスへのアクセス記録 (ログデータ) を解析するのが Web ログマイニングである。Web ログマイニングには、そのサービスのユーザビリティ向上や、さらには該当分野の知識 (マーケティングデータ) が獲得できることへの期待がもたれている。

一方、現在注目を集めている情報提供サービスの一つが、位置にもとづく情報提供サービス (本稿では位置情報サービスと呼ぶ) である。位置情報サービスは、従来から地図やお店探しなどの日常生活に応じた検索サービスとして親しまれてきたが、現在では GPS 搭載の携帯電話

普及などによる現地周辺の提供サービスや、サーチエンジンの分野におけるローカルサーチというコンセプトも導入され、より身近なものになってきている。

本研究の対象である位置指向の Web ログマイニングは、「地理ログデータ」、すなわち地理的位置に関連した情報へのアクセス記録を大量に解析して、そこから有意義な知識を得ることである。研究の課題は大きく次の2つになる。

1. ある場所に何があるかをログデータから明らかにすること (場所の特徴)
2. ある「もの」はどこにあるかをログデータから明らかにすること (事物の地理的特徴)

本研究はインターネット上の位置情報サービスの大量検索ログデータを解析しながら上

記課題にアプローチする。今回対象とするログデータは、「住所ワード」と「フリーワード」の入力により情報検索を提供するサービスのものであり、解析するログデータは、利用者が入力した「住所ワード」と「フリーワード」の組が時系列に並んだものである。位置情報サービスの位置指定は、他にデバイスで位置を測位する方式やメニューで場所を指定するものがあるが、このキーワード方式は現状では最も親しまれているものである。このデータから抽出可能な相関ルールは主に次のようになる。

- ・住所⇒住所
- ・フリーワード⇒フリーワード
- ・住所⇒フリーワード
- ・フリーワード⇒住所

本稿では、これらルールの解析法について次の構成で報告する。まず2章では本研究で利用したログデータの概要を説明する。さらに3章では相関ルール解析の方法について説明し、4章では解析したルールから実用上「興味深い」ルールを取り出す方法について考察する。5章で、取り出したルールを検索システムのフロントエンドに適用したキーワード推薦システムを紹介し、6章で関連研究に触れた上、7章で結ぶ。

2. 解析対象データ

2.1. データの定義

本実験で解析に利用したデータはタウン情報検索のデータである。表1に例を示す。

表1 実験に利用したログデータ

時間	住所ワード	フリーワード	cookie
20/Sep/2003:08:19:48	東京都渋谷区	ホテル	xxxxxxx
20/Sep/2003:08:20:30	東京都港区	レストラン	xxxxxxx

ログデータの単位行をリクエストと呼ぶ。リクエストは、利用者が入力した住所ワードとフリーワードを含むタイムスタンプ付きのデータである。このリクエストをまとめたものをセッションとして解析に用いる。

2.2. データ処理の概要

データ処理の概要を示す。ログデータは初めにリクエスト単位に解析に必要な項目を抜粋した後、極端なリクエストを削除しながら、セッションに分割する（以上の作業をクリーニングと呼ぶ）。具体的には以下のとおりである。

- ・項目抜粋は、リクエストを解析して前節で示した表1の形式を取り出した。なお、住所ワードとフリーワード両者に入力があるリクエストのみを採用した。
- ・次に同一利用者からの連続的なリクエストまとめてセッションとする。本実験では、同一ユーザが前後30分を超えない間隔で行ったリクエストを同一セッションに属すると定義した。ユーザの同一性に関してはhttp-cookieを用いた。

セッション化にあたって、crawler（データ収集目的の大量リクエスト）や、試用ユーザからのリクエストを解析から除外する目的で極端なリクエストの削除を行った。極端であると判断したのは、セッション長（1セッションで行った検索リクエスト回数）が40を超えるもの、および対象全データ中で、リクエスト回数が1回だけのもの、およびリクエスト回数が、500回を超えるものである。

2.3. データの概要

クリーニングを施した後の解析対象データの概要を図1に示す。全体を3期間に分けて累計で示している。全体で約160万のユーザからの1,400万のリクエストが、500万のセッションに分割されている。ユニークな住所ワードは42万、フリーワードは240万程度である（ただしこの単語には、利用者の操作ミスによる、非単語入力や文字化け等で解析できない文字列も含んでいる）。データサイズはオリジナルデータが約150ギガバイト、クリーニング後が約800メガバイトであった。

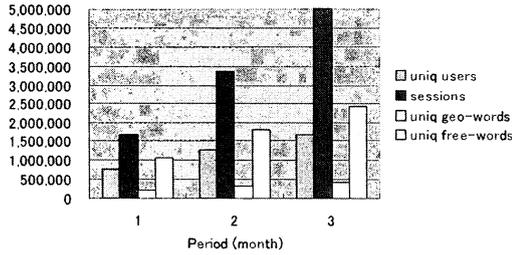


図1 データの概要

表2 データの概要

uniq users	sessions	uniq geo-words	uniq free-words	accesses
756,267	1,676,145	208,454	1,055,147	4,766,618
1,259,096	3,360,972	329,399	1,799,566	9,528,445
1,678,474	5,032,465	428,747	2,437,078	14,196,016

3. 解析手法

今回作成したセッションから次の関係が定義できる。

- ・セッションベースのルール
- ・リクエストベースのルール

セッションベースのルールは、セッションを「バスケット」、キーワードを「購入アイテム」とみなした時のルールである。セッションは入力された住所ワード (geo-word; gw)、フリーワード (free-word; fw) から構成される。

この中から、 $gw \Rightarrow fw$ 、 $fw \Rightarrow gw$ 、 $gw \Rightarrow gw$ などのルールが得られる。すなわち、アイテム集合を $I = \{gw_1, fw_1, gw_2, fw_2, \dots, gw_n, fw_n\}$ 、 $X, Y \subset I$ 、としたときの相関ルール $X \Rightarrow Y$ に関して、 N を総セッション数、 $N(X)$ を X を含むセッションの数とすると、 $X \Rightarrow Y$ の支持度 sup、確信度 conf、 $X \Rightarrow Y$ 間の相互情報量 Info、はそれぞれ次のようになる。

$$\begin{aligned} \text{sup}(X \Rightarrow Y) &= N(X \cup Y) / N \\ \text{conf}(X \Rightarrow Y) &= N(X \cup Y) / N(X) \\ \text{Info}(X:Y) &= \log_2 \{N(X \cup Y) * N / (N(X) * N(Y))\} \end{aligned}$$

また、本データは常に (gw, fw) の組でリクエストが行われているという特徴があるため、リクエストベースのルールの定義も可能である。この場合 N は、総リクエスト数、 $N(X \cup Y) = N(gw, fw)$ となる。なおこのとき、fw を gw を発生させ

る情報源と見なすとエントロピー H は、以下のようになる。

$$H(fw) = - \sum (N(gw, fw) \log (N(gw, fw) / N(fw)))$$

4. 相関ルール抽出実験と考察

3章で説明したログデータから相関ルールを抽出した。最小支持度 0.000002 の元で約 270 万通りのルールが作成された。本ルールに特徴的な点は、アイテムが住所ワードであるか、フリーワードであるのかがあらかじめわかっている点である。この観点からルールは次のように分類できる。

- ・フリーワード \Rightarrow 住所ワード
- ・住所ワード \Rightarrow フリーワード
- ・住所ワード \Rightarrow 住所ワード
- ・フリーワード \Rightarrow フリーワード
- ・その他

この分類を踏まえて、取り出されたルールの数と例を表3に示す。

表3 抽出されたルールの例

分類	ルール数	例
住所の2項ルール	187,534	東京都目黒区駒場 \Rightarrow 東京都渋谷区
フリーワードの2項ルール	79,884	出前 \Rightarrow ラーメン
住所フリーワード間の2項ルール	276,130	愛媛県松山市 \Rightarrow 道後温泉
住所の3項ルール	1,850,256	道頓堀 \Rightarrow 心齋橋筋 and 千日前
フリーワードの3項ルール	184,212	ラーメン \Rightarrow 焼肉 and 中華料理
住所フリーワード間の3項ルール	175,836	旭屋書店 and 大阪市 \Rightarrow ジュンク堂

4.1. フリーワードと住所ワードの関係

フリーワードから住所へのルールが求まると、検索システムにおいてふさわしい住所を推薦することが期待できる。ここで抽出されたルールの適用性について考察する。例としてフリーワードが「ホテル」であるときの $fw \Rightarrow gw$ ルールを考える。「ホテル」を左項に持つルールはリクエストベース解析だけでも 1000 種類を超えるため、不要なルールの削除方法またはルールのランク付け方法が必要になる。表3に「ホテル」の典型的なルールについて、支持度と相互情報量とともに示した。相互情報量 (もしくは Lift 値) はルールに correlation を求めるため

の指標として知られている。またルールの左項が同一のときは、支持度と確信度は本質的に同一なので確信度は省略している。

表 4-a は最小支持度を高く設定しているが、その結果、ルールに現れる gw の出現確率は任意の fw に対する gw とほぼ同等になってしまい、高い意味を見出しにくい。一方、最小支持度を 1/10 下げると、比較的観光地として知られた住所が現れてくる (表 4-b)。さらに低い支持度では、住所の粒度も小さく、特定の地域が提示される (表 4-c)。

表 4 ホテルに関するフリーワード住所間ルールの例
各支持度のクラス中で相互情報量の高いものを提示
(上より順に表 4-a、b、c とする)

最小支持度>0.0001		支持度	相互情報量
ホテル ⇒	東京都千代田区	0.000142	0.2
ホテル ⇒	京都府京都市	0.000134	0.1
ホテル ⇒	東京都港区	0.000132	0.0
ホテル ⇒	東京都新宿区	0.000134	0.0
ホテル ⇒	北海道札幌市	0.000108	-0.3

最小支持度>0.00001		支持度	相互情報量
ホテル ⇒	千葉県浦安市舞浜	0.000015	4.1
ホテル ⇒	新大阪(JR東海道新幹線/大阪府)	0.000011	3.5
ホテル ⇒	鹿児島県指宿市	0.000013	3.0
ホテル ⇒	新横浜	0.000011	2.7
ホテル ⇒	神奈川県足柄下郡箱根町	0.000018	2.5

最小支持度>0.000001		支持度	相互情報量
ホテル ⇒	宮城県宮崎市高千穂通	0.000002	4.5
ホテル ⇒	舞浜(JR京葉線/千葉県)	0.000003	4.4
ホテル ⇒	沖縄県恩納村	0.000003	4.3
ホテル ⇒	静岡県浜松市舘山寺町	0.000002	4.2
ホテル ⇒	北海道虻田郡洞爺村	0.000003	4.2

一般に検索システムにおける住所入力における補助の必要性としては、

- ・ 代案の提示 (近くの代替りの住所)
- ・ 最適案の提示 (名産地)
- ・ タイプミスの修正

などが考えられる。「ホテル」などの分析から、関連する住所には複数のコンテキストがありう

ること、および、特に相互情報量などの correlation を示す指標を用いて相関ルールを分類することにより、一般的な地名と名産地や活動の中心地である住所を分けて提示することが可能であると考えられる。

次に図 2 にフリーワードのエントロピーを計算した結果を示す。

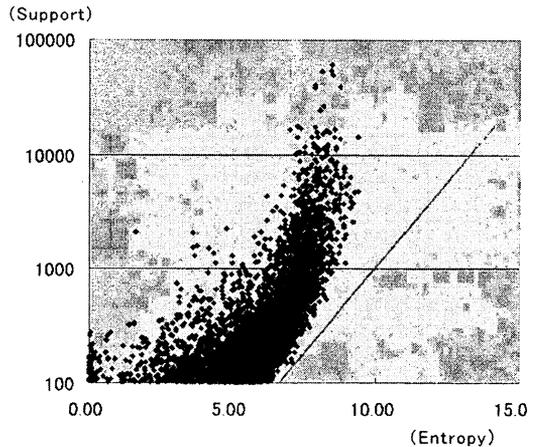


図 2 入力フリーワードのエントロピー
x 軸をエントロピー、y 軸を支持度 (出現回数) とする。右の直線はそれぞれの支持度におけるエントロピーの理論限界を示す。

このエントロピー値は、ある「フリーキーワード」で住所が「どの程度決まるか」を示す値である。支持度が高いワード群にはエントロピーの大きな差はなく、実際にどこにでもある職業・サービス (commodity) に対応するようなワードがしめている (例: 農協、旅館、官公庁、消防、役場、中学校)。一方、支持度が低いワード群は、エントロピーに比較的開きがある。エントロピーがより高いグループには commodity に比べて、店舗数等は必ずしも多くはないが、親しまれている業種が (例: 熱帯魚店、小僧寿し、パソコンワープロ教室、森林組合)、低いグループには専門的なサービスが (発電機、調味料製造業、鉄砲店、邦楽教授所、) 位置づけられている。なお、エントロピーをさらに低くすると、固有名詞が多くなる。「渋谷駅」や「大阪ドーム」といったフリーワードは、理論通りに 0 または 0 に近いエントロピーを持つ。

この分析から、ワードをエントロピーと支持

度（出現頻度）を用いて、commodity であるか否か、地域特性があるか否か、という予測を立てることが考えられる。この予測を使えば、commodity に対しては、比較的近隣のよくある住所を、専門的なワードに対しては、特産・名産地、もしくは分野の中心的な住所を提示することができる。

4.2. 住所ワード間の関係

次に住所ワード間の関係を与えるルールについて考察する。3 章でのべた方法で、セッション分析を行い地名ワード間の関係を求めた。表 5 に得られたルールの分類を、表 6 にその例を示す。表 5 から特に市区町村から市区町村へのルールが多種（71,612 種）抽出されている。次に表 6 はルールの例を示したものである。この例はそれぞれの地名に関して、ルールを支持度の高いものから 5 件ずつ取り上げた。この例のように、支持度の高いルールは、(1)上位地域（都道府県）、(2)近接地域、を主として含んでいる。また「所沢市」のケースではその生活圏である「東京」がルールとして現れている。本例に以外のデータから、多くの例において、低い支持度においては、より細かい粒度の住所が現れてくる傾向がある。すなわち、単純な支持度によるルールの使用が、（前節で述べた）commodity に対する近接の住所の推薦に役に立つことが考えられる。

表 5 抽出された住所間の相関ルール。
得られた A⇒B なるルールのユニーク数を示す。

A \ B	種類数	都道府県	市区町村	町大字	字丁目	その他	合計
都道府県	47	2,162	9,262	2,419	426	3,673	17,942
市区町村	2,462	9,262	71,612	11,686	1,112	10,359	104,031
町大字	5,740	2,419	11,686	13,170	2,395	2,362	32,032
字丁目	2,178	426	1,112	2,395	2,186	355	6,474
その他	5,690	3,673	10,359	2,362	355	10,306	27,055

表 6 住所ワード間の相関ルール

		支持度	相互情報量
東京都港区六本木	⇒ 東京都港区	0.00027	4.9
東京都港区六本木	⇒ 東京都	0.00013	1.0
東京都港区六本木	⇒ 東京都港区赤坂	0.00004	5.5
東京都港区六本木	⇒ 東京都渋谷区	0.00003	1.9
東京都港区六本木	⇒ 東京都港区西麻布	0.00003	6.9
埼玉県所沢市	⇒ 埼玉県	0.00045	3.7
埼玉県所沢市	⇒ 東京都	0.00014	-0.3
埼玉県所沢市	⇒ 埼玉県狭山市	0.00012	6.1
埼玉県所沢市	⇒ 埼玉県川越市	0.00011	4.6
埼玉県所沢市	⇒ 埼玉県入間市	0.00011	6.2

4.3. 住所の表記補正と階層を使ったルールの集約

本研究の対象のように対象アイテムの空間がスパースな場合は、ルール作成、およびルールの応用時において、ルールに優先順位をつけること、およびルール自体の数を減らし、有意なルールを見つけることが必要である。ルールの優先順位に関する考え方は、前節までで述べたが、本節では、ルールの数に関するケーススタディーを紹介する。表 7 は住所間ルールに関する分析である。本分析で対象となった住所ワードは、入力されたままの文字列形式で 18,028 種であり、そこから作成されるルールは約 20 万種であった。これに住所の表記補正をかけると、住所ワードが 1,600 種に、また市区町村レベル以下の細かいレベルの住所を、市区町村レベルにまとめると、住所ワードは 8,000 種になることがわかった。表記の統一はもともとの情報量を保ったままのルールの集約が可能であるし、階層を使った集約も、例えば市区町村レベルに統一すれば、検索サービスにおいて十分実用的であるケースが存在するので、相関ルールを使うために効果的である。なお、ここで行った住所の表記補正は、主に次の通りである。

- ・ 住所 suffix の省略補完（神奈川⇒神奈川県）
- ・ 上位住所の省略補完（横浜市⇒神奈川県横浜市）
- ・ 数字表記の統一（四丁目⇒4丁目）

表 7 住所の表記補正と階層を使ったルールの集約

	オリジナル	表記補正	市区町村レベル
ユニーク住所ワード数	18,028	16,117	8,199
ユニークルール数	208,784	187,534	132,637

5. 位置にもとづく検索サービスのためのキーワード推薦システム

前章までで、作成したルールを使ったキーワード推薦プロトタイプシステムを作成した。図 4 に示す。

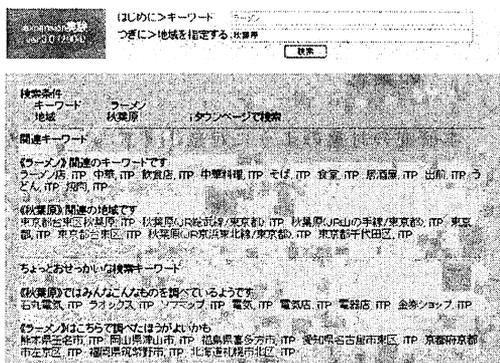


図 4 キーワード推薦システム

このシステムはタウン情報検索サービスのフロントエンドとして独立に構築したもので、ユーザーのリクエストを受け取り、それに「ふさわしい」単語を推薦する。横にある iTP アンカーをクリックすることにより、推薦結果に変更した検索をタウン情報システムに対して行うことができ、また変更結果をもとに単語間の関係を見ながら、ブラウズを続けることもできる。図示した例では、「ラーメン、秋葉原」という検索を行っている。これに対して、システムは画面の上半分「秋葉原」の正式住所を含む近隣等の関連住所を提示する。また同時に、「ラーメン」を specialty と考えたときの住所が下半分に表示される。これらはそれぞれ順に 4.2 節、4.1 節で述べた方法によるルールを出力している。また本論文では述べなかったが、住所ワード⇒フリーワードの関係の逆で、フリーワード⇒住所ワードに関する同様の分析も可能である。その分析結果を使って、フリーワードの推薦も行って

いる。specialty 推薦では、秋葉原をキーに、電気街に関する関連ワードが提示できている。

6. 関連研究

本研究で使っている相関ルールは Agrawal らの提案[1]にもとづくものである。相関ルールの問題点は、計算量に関するものと、利用上の問題に分けられ、後者に関するものには、ルールの量（優先順位や数の抑制）、およびルールの質（興味深いルールを取り出すこと）の問題がある。ルールの「興味深さ」に関しての研究はいくつか知られている[2][3]。本研究は、興味深いルールを取り出す指標として知られている相互情報量（もしくは Lift 値）[4]を使うことをベースに位置に関する考察を行っている。

地理的な統計の解析は従来からマーケティングリサーチ分野で行われてきているものであるが、サーチエンジンのログ情報を元に位置指向に解析する研究は必ずしも多くない。筆者らの研究[5]では比較的粒度の大きいサービスを対象に分析を行ったが、本研究は対象とするアイテム数（住所ワードとフリーワード）が約 300 万とより大規模になっている。

キーワード推薦の研究は、従来から特に collaborative filtering として数多くの研究がされている。筆者らの研究[6]では職業分類をクラスタリングして、キーワード推薦を行う提案をしているが、これは統制語を対象語彙とした(約 2000 種)。本研究は、対象を住所にも広げ、より大規模な語彙を対象にしている。

7. おわりに

Web ログデータを、ユーザーのリクエストの地理的な性質に着目してマイニングを行う手法およびそのケーススタディーに関する報告を行った。

解析は約 150 ギガバイトのログデータから、約 160 万のユーザ、住所ワード 42 万通り、フリーワード 240 万通りのデータにクリーニングし、そこから最小支持度 0.000002 の元で約 270 万通りの位置ワードと位置に関連したフリーワードの相関ルールを作成した。

解析結果から、住所ワードの推薦には、commodity である情報対象に対する近隣地域の提示、および specialty に対する中心地域の推薦が可能であること、およびそれぞれのルールを選ぶ際には相関ルールの支持度だけでなく、単語間の相互情報量が有効であることを述べた。また本解析結果を使った、単語を推薦して検索の補助を行うシステムについて説明した。

今後は本ルールおよび推薦システムの評価、およびルールの可視化を行う予定である。

謝辞

本研究にご協力いただいた NTT 番号情報株式会社に感謝します。

文 献

- [1] Rakesh Agrawal, Tomasz Imielinski, Arun N. Swami, "Mining Association Rules between Sets of Items in Large Databases". SIGMOD Conference pp.207-216. 1993
- [2] Bayardo R.J. and R. Agrawal, "Mining the Most Interesting Rules". Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, pp. 145-154. 1999
- [3] Pang-Ning Tan, Vipin Kumar, "Interestingness Measures for Association Patterns : A Perspective", KDD'2000 Workshop on Postprocessing in Machine Learning and Data Mining, 2000
- [4] 岡田孝、元田浩: "相関ルールとその周辺", オペレーションズ・リサーチ, vol.47, no.9, pp.565-571. 2002
- [5] Iko Pramudiono, Takahiko Shintani, Katsumi Takahashi, Masaru Kitsuregawa, "User Behavior Analysis of Location Aware Search Engine", Proc. of International Conference On Mobile Data Management (MDM'02), IEEE Computer Society Press, pp. 139-145. 2002
- [6] Yusuke Ohura, Katsumi Takahashi, Iko Pramudiono, Masaru Kitsuregawa, "Experiments on Query Expansion for Internet Yellow Page Services Using Web Log Mining", VLDB 2002: pp.1008-1018. 2002