

女性向けフリーマガジン発行サイトにおける 店舗ページ推薦に関する検討

高橋修太郎[†] 大塚 真吾^{††} 宮崎 収[†]

[†] 千葉工業大学 情報科学部 〒 275-0016 千葉県習志野市津田沼 2-17-1

^{††} 独立行政法人 物質・材料研究機構 〒 305-0047 茨城県つくば市千現 1-2-1

あらまし 本稿では都内で働く女性を対象としたフリーマガジンと連動するサイトにおける店舗ページの推薦方法に関して検討を行った。このサイトにアクセスするユーザは都内で働く女性が多いという特徴があり、アクセスログデータの解析を行った結果、リラクゼーションと同時に他の異なるジャンルに属する店舗ページを閲覧するユーザが多いなど、興味深い知見が得られた。

キーワード アクセスログ解析, Web ページ推薦, Web 行動調査, Web 空間と実世界

A Study for Shop Recommendation in a Free Magazine Site for Women

Syutaro TAKAHASHI[†], Shingo OTSUKA^{††}, and Nobuyoshi MIYAZAKI[†]

[†] Department of Computer Science, Faculty of Information and Computer Science, Chiba Institute of Technology
Tsudanuma 2-17-1, Narashino-shi, Chiba, 275-0016 Japan

^{††} Scientific Information Office, National Institute for Materials Science
Sengen 1-2-1, Tsukuba-shi, Tsukuba, 305-0047 Japan

Abstract We study for shop recommendation in a web site which is connected free magazine for working women in Tokyo. There is characteristic that many users accessed this site are working women in Tokyo. Some interesting facts are found in the results of access log analysis. For example, there are many users who viewed shopping pages belonging to relaxation genre and other different genres simultaneously.

Key words Web access logs analysis, Recommendation of Web pages, Investigation of behavior on the web, Relation between web and real world

1. はじめに

インターネットの爆発的な普及により、自分の興味がある情報の探索、購入予定の商品やレストラン、旅館の評判探索など、ユーザが望む情報を Web 空間から手軽に入手できるようになった。

Web 空間におけるユーザ行動を支援するための研究は盛んに行われており、例えば、ユーザの行動抽出、ある分野に関する評判の自動抽出、検索支援、商品の推薦などあらゆる研究が行われている。さらに、あるサイトにおけるユーザの行動分析を詳細に解析することで、Web ページのアクセス数向上や Web 広告について統合的にサポートするビジネスも誕生している。

ユーザの行動分析をするためにはユーザの特性を把握する必要があるが、既存の研究では、

- ある特定の内容に特化したサイトのアクセスログ
- 検索エンジンやポータルサイトなど不特定多数のユーザがアクセスを行うサイトのアクセスログ

- 各検索エンジンが提供している検索ランキングやスニペットなど検索結果に関連する情報

- 一般に公開されている大学の proxy サーバや NASA などのアクセスログなどを利用しており、Web 空間におけるユーザの行動に限定したものである。

アクセスログを用いた研究は、例えば、Web ページやユーザのクラスタリング [2], [13]、ユーザの行動抽出 [4], [11]、検索支援 [1], [5], [10], [12]、閲覧行動支援 [6] など数多くの研究が行われている。これらの研究は主に Web 空間におけるユーザの行動支援を目的としているため、実世界の行動との関連に関しては議論されていない。実世界と Web 空間の両者に関連する研究としては、宇陀らの図書館サービスの利用者を対象とした利用状況の調査 [9] や、柴田らのライフログを用いた子供の事故防止対策に関する研究 [7] などが行われている。

また、上記の研究で用いられている主なログデータとしては

- 自サイトまたは共同研究している企業のアクセスログ



図1 フリーマガジン「Well」のトップページ

データ

- The Internet Traffic Archive で公開されているアクセスログデータ [8]
- インターネット視聴率調査会社が提供するウェブサイト視聴データ (パネルログ)^(注1)
- Overture が提供するクエリログ (キーワードアドバイスツール)^(注2)

などがあるが、基本的には不特定多数のユーザを対象にしたデータのため、我々が用いた年齢層や性別が偏ったアクセスログとはその性質が異なる。

我々は、以前、都内の女性 OL をターゲットとしたフリーマガジンと連動するユーザ属性がある程度定まったサイトのアクセスログを用いて、アクセス時間や検索語に焦点を当てたユーザの行動解析を行った [3]。本稿では、ユーザに対して店舗ページの推薦を行うための予備解析を行い、そこから得られた新たな知見を示す。

2. フリーマガジンとアクセスログデータの詳細

まず、今回我々が解析を行ったサイト^(注3)と連動しているフリーマガジン「Well」(図1)についての説明を行う。この雑誌は偶数月の20日に発行されており、以下の方法で毎月約30万部が配布されている。

- 朝と夕方に、丸の内、銀座、六本木などで街頭配布 (発行日から一週間程度)
- 地下鉄の駅に設置されているラック (約20箇所)
- コンビニに設置されているラック (約750箇所)
- 企業配布 千代田・中央・港区約4,000オフィス (約7万人定期読者)
- 都内飲食店、スクール、美容院、など設置協力店 (約500

(注1): ネットレイティングスやビデオリサーチインタラクティブが有償で提供するデータ

(注2): 2007年5月に正式サポートを終了

(注3): <http://my-well.net>

店舗)

雑誌の内容は銀座・丸の内・青山・六本木など山手線の内側をメインターゲットにしたものになっており、雑誌で紹介されている店舗は以下の8つに分類されている。

- (1) グルメ&ドリンク
- (2) ビューティー
- (3) アミューズメント&レジャー
- (4) キャリア&ライフ
- (5) ショッピング&ピックアップ
- (6) ヘルス&クリニック
- (7) ヘア&メイク
- (8) リラクゼーション

この雑誌に掲載された店舗情報はすぐに連動するサイトに掲載される仕組みとなっている。このサイトでは今のところユーザ登録などの仕組みを導入していないため、アクセスしたユーザの年齢や男女比などを正確に掴むことはできないが、サイト内で使われた検索語やサイトに流入するために使われた検索語の調査や読者アンケートから20-30代の女性が大多数を占めていると考えられる。

今回解析を行ったアクセスログは2004年2月から2008年12月まで^(注4)であり、データ量は約8Gbyteであった。ここから、html,htm,phpのファイルのみを抽出し、さらにWebクローラーなど自動巡回ツールからのアクセスを取り除くクリーニング処理を行うことで、最終的には250万アクセスとなった。また、アクセス元のIP情報の種類は約55万であった。1人で複数のIPアドレスを使用していたり、アクセス毎にIPアドレスが変更されることがあるため、ユーザ数の詳細については分からないが、少なくとも数千から数万人のアクセスがあったと考えている。また、クリーニングしたログデータから、IPアドレスを元に約87万セッションを抽出し解析を行った。^(注5)

今回解析を行ったアクセスログはApacheサーバのログのため、アクセスしたユーザのIPアドレス、アクセス時間、アクセスされたページ、リファラー、ブラウザなどの情報を含んでいる。また、各店舗はWellコードと呼んでいるIDを割当てており、サイト内の各店舗ページはこの番号を元に自動生成される仕組みになっている。Wellコードからジャンルを特定することが可能なため、今回の解析では店舗ページのアクセス以外にも、ジャンル毎にアクセス状況を調べた。

3. 解析結果

アクセスログから得られる情報を元に今回の解析では以下の調査を行った。

- (1) 各店舗ページのアクセス数の傾向
- (2) 同一セッション内で閲覧された店舗ページの傾向
- (3) 同一セッション内で閲覧されたジャンルの傾向

(注4): 途中、2005年1月-5月と2006年7月-12月はデータが欠落している。

(注5): セッションの抽出方法: 1. アクセスログを同一IPでまとめ、アクセス時間順にソートする。2. アクセス時間が30分以上開いた場合はセッションの分割を行う。

表 2 同一セッション内で閲覧された店舗ページ

グルメ&ドリンク	23,189							
ビューティー	1,565	5,934						
アミューズメント&レジャー	2,285	437	4,758					
キャリア&ライフ	1,620	571	732	2,676				
ショッピング&ピックアップ	1,389	577	329	1,146	3,353			
ヘルス&クリニック	822	189	119	222	283	1,216		
ヘア&メイク	617	349	85	205	137	40	1,293	
リラクゼーション	3,923	2,264	1,847	819	1,261	292	292	10,438
	グルメ&ドリンク	ビューティー	アミューズメント&レジャー	キャリア&ライフ	ショッピング&ピックアップ	ヘルス&クリニック	ヘア&メイク	リラクゼーション

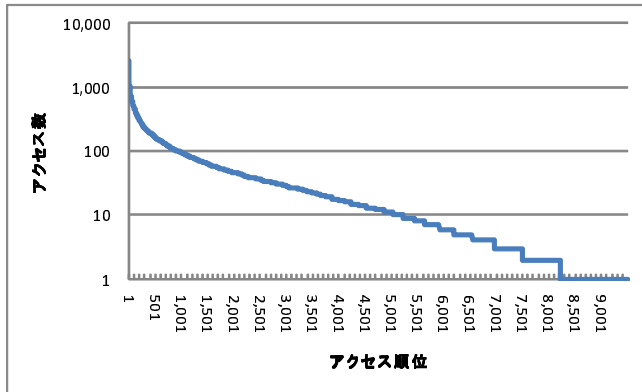


図 2 店舗ページのアクセス数の傾向

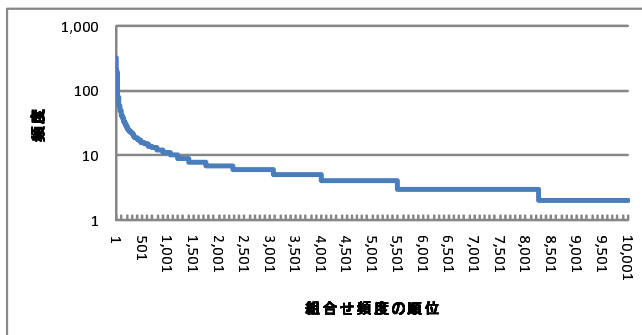


図 3 同一セッション内で閲覧された共起する店舗ページのアクセス傾向

3.1 店舗ページのアクセス数の傾向

まず始めに店舗ページのアクセス数の調査を行った。店舗ページのアクセス数の合計は 394,679 であった。店舗ページのアクセス件数が全セッション数（約 87 万セッション）より少ないことから、店舗ページにアクセスしないケースが数多く存在し、ユーザに対する店舗ページ推薦の必要性を表している。

アクセスが多い順に店舗をソートした結果を図 2 に示す。図の横軸はアクセスが多い順に店舗ページを並べた時の順位であり、縦軸はその店舗ページに対するアクセス数を示している。なお、縦軸については対数としている。結果から、アクセス数はべき乗則となっていることが分かる。

3.2 同一セッション内で閲覧された店舗ページの傾向

複数の店舗ページを閲覧したセッションを対象に、ユーザが同一セッション内で閲覧した店舗ページの組合せの調査を行った。ユーザが閲覧した店舗ページの組合せ数は約 49,000 であり、その頻度の合計は 105,792 であった。頻度の傾向については図 3 に、その組合せの内容については表 1 に示す。図 3 の横軸は組合せ頻度が多い順に並べた時の順位であり、縦軸は頻度

表 1 同一セッション内で閲覧された共起する店舗ページのジャンル

組合せ頻度の順位	ジャンル
1	グルメ&ドリンク グルメ&ドリンク
2	グルメ&ドリンク グルメ&ドリンク
3	グルメ&ドリンク グルメ&ドリンク
4	グルメ&ドリンク グルメ&ドリンク
5	グルメ&ドリンク グルメ&ドリンク
6	ヘルス&クリニック グルメ&ドリンク
7	グルメ&ドリンク グルメ&ドリンク
8	グルメ&ドリンク グルメ&ドリンク
9	グルメ&ドリンク グルメ&ドリンク
10	ヘア&メイク ヘア&メイク
11	ヘア&メイク ヘア&メイク
12	グルメ&ドリンク グルメ&ドリンク
13	ビューティー ビューティー
14	ビューティー ビューティー
15	ビューティー ビューティー
16	ビューティー ビューティー
17	ビューティー ビューティー
18	ビューティー ビューティー
19	ビューティー ビューティー
20	ビューティー ビューティー
21	ビューティー ビューティー
22	ビューティー ビューティー
23	ビューティー ビューティー
24	ビューティー ビューティー
25	ビューティー ビューティー
26	ビューティー ビューティー
27	ビューティー ビューティー
28	ビューティー ビューティー
29	ビューティー ビューティー
30	ビューティー ビューティー

を示している。なお、縦軸については対数としている。この結果についても、べき乗則となっていることが分かる。

また、表 1 については、組合せ頻度が多い上位 30 件について、店舗ページが属するジャンルを示したものである。結果から、6 位と 17 位以外については同一ジャンルであることが分かる。また、グルメ&ドリンクとヘア&メイクど同士の組合せが多いことが分かる。上位 100 件まで調査した結果、ジャンルが異なった組合せは 3 件だけであった。一方、順位が下位の組合せ 100 件についてランダムに抽出して同様な調査を行ったところ、ジャンルが異なった組合せは 27 件であった。

3.3 同一セッション内で閲覧されたジャンルの傾向

最後に、同一セッション内で閲覧した店舗ページの組合せについて、そのページが属するジャンルの調査を行った。その結果を表 2 に示す。表中のジャンル名は店舗ページが属するジャンルを示しており、例えば、左軸のリラクゼーションと縦軸のビューティーが交差する 2,264 という値は、同一セッション内でリラクゼーションに属する店舗ページとビューティーに属する店舗ページを同時に見たセッション数が 2,264 あったことを示している。

この表から、同一ジャンルへの移動が多く、それ以外ではリラクゼーションに関するページは他のジャンルに属する店舗ページとの組合せが多いことが分かる。また、キャリア&ライフとショッピングピックアップの組合せや、グルメ&ドリンクとアミューズメント&レジャーの組合せも多いことが分かる。

4. おわりに

本稿では、都内の女性 OL をターゲットとしたフリーマガジンと連動するユーザ属性がある程度定まったサイトのアクセスログを用いて、ユーザに対して店舗ページの推薦を行うための予備解析を行い、店舗ページ推薦の可能性を模索した。

解析結果から、ユーザは同一のジャンルに属する店舗ページが多いことや、リラクゼーションに関するページは他のジャンルに属する店舗ページと一緒に閲覧されているといった知見を得ることができた。今後はページ推薦を実際に行うためにより詳細な調査と、時間や曜日によるユーザアクセスの傾向からのアプローチについても検討を行う予定である。

謝辞 本研究の一部は科学研究費補助金 基盤研究 (C)(課題番号 21500107) の助成による。また、本研究を進めるにあたってアクセスログの提供をして頂いた、株式会社 ウェル 白神志郎様に深謝致します。

文 献

- [1] D. Beeferman and A. Berger. Agglomerative clustering of search engine query log. In *The 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2000)*, August 2000.
- [2] T. Murata. Discovery of user communities from web audience measurement data. In *The 2004 IEEE/WIC/ACM International Conference on Web Intelligence (WI2004)*, September 2004.
- [3] 大塚真吾, 高久雅生, 喜連川優, 宮崎収兄. 女性向けフリーマガジン発行サイトにおけるユーザの行動分析. *DEIM2009*, pp. 8B-4, 2009/3.
- [4] 大塚真吾, 豊田正史, 喜連川優. ウェブコミュニティを用いた大域 web アクセスログ解析法の一提案. *情報処理学会論文誌: データベース*, Vol. 44, No. SIG18(TOD20), pp. 32-44, 2003/12.
- [5] 大塚真吾, 豊田正史, 喜連川優. 大域ウェブアクセスログを用いた関連語の発見法に関する一考察. *情報処理学会論文誌: データベース*, Vol. 46, No. SIG18(TOD26), pp. 82-92, 2005/6.
- [6] 齋藤皓太, 村田剛志. Web 閲覧者の関心キーワードの抽出と巡回行動の可視化. *電子情報通信学会 WI2 研究会資料*, pp. 141-146, 2006/3.
- [7] 柴田康徳, 本村陽一, 西田 佳史 山中龍宏, 溝口博. 日常モノデータベースとライフログとの統合による危険の可視化. 第 21 回人工知能学会全国大会, 2007/6.
- [8] The internet traffic archive . <http://ita.ee.lbl.gov/>.
- [9] 宇陀則彦, 伊藤宏美, 松村敦. アクセスログに見る電子図書館利用の傾向. *情報知識学会誌*, Vol. 18, No. 2, pp. 161-168, 2008/9.
- [10] J. Wen, J. Nie, and H. Zhang. Query clustering using user logs. *ACM Transactions on Information Systems (ACM TOIS)*, Vol. 20, No. 1, pp. 59-81, January 2002.
- [11] 山田和明, 中小路久美子, 上田完次. Web ユーザの行動履歴解析のためのデータマイニング. *電子情報通信学会 WI2 研究会資料*, pp. 59-64, 2005/9.
- [12] 安川美智子, 横尾英俊. クエリログから獲得した関連語のクラスタリングに基づく web 検索. *電子情報通信学会論文誌 D*, Vol. J90-D, No. 2, pp. 269-280, 2007/2.
- [13] H. Zeng, Z. Chen, and W. Ma. A unified framework for clustering heterogeneous web objects. In *The Third International Conference on Web Information Systems Engineering (WISE'02)*, December 2002.