様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年　6 月　1 日現在

機関番号：３２５０３
研究種目：基盤研究（Ｃ）
研究期間：2009～2011
課題番号：２１５００１０７
研究課題名（和文）　ウェブ空間に対する紙メディアの影響力分析に関する研究

研究課題名（英文）　The influence of paper media on the Web surfing behavior

研究代表者
宮崎　収兄　（MIYAZAKI NOBUYOSHI）
千葉工業大学・情報科学部・教授
研究者番号：２０２６５４６６

研究成果の概要（和文）：フリーペーパーやテレビなど、身近なメディアからの情報をもとにインターネットへアクセスを行うユーザの行動解析や分析を行った。その結果から、ウェブ空間と実世界を動く人々の行動パターンを明らかにし、ウェブ空間における紙メディアなど実世界の影響力を分析することができた。

研究成果の概要（英文）：When magazines are published or TV programs are broadcasted, people sometimes access the Internet to look at related Web sites or Blogs. We study the relationship between these media in the real world and the peoples' behavior in the Internet. We analyzed access logs of a Web site operated by a free magazine and various Blog contents that describe and discuss TV programs. This report discusses the influence of such media on the Web surfing behavior.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,600,000	480,000	2,080,000
2010 年度	1,100,000	330,000	1,430,000
2011 年度	800,000	240,000	1,040,000
年度			
年度			
総　計	3,500,000	1,050,000	4,550,000

研究分野：総合領域
科研費の分科・細目：情報学・メディア情報学・データベース
キーワード：Web サービス、アクセスログマイニング、ウェブと実社会

１．研究開始当初の背景

インターネットに誰でも気軽にアクセスできるようになり、人々のライフスタイルも様変わりしてきた。例えば、Amazon や楽天に代表されるネットショッピングやYahoo!に代表されるネットオークションなど、インターネットが普及する前には考えられないサービスがごく当たり前に提供され、人々はウェブ上でも商品の購入が手軽に行えるようになった。ウェブ空間におけるユーザの行動解析・分析に関する研究は、大きく分けて以下の３つのタイプに分類される。

1) ウェブサーバのアクセスログを用いて、同一サイト内におけるユーザの行動解析や購買パターンの分析を行う研究。

2) 大学のプロキシーサーバのログやブラウザのログを用いて、ウェブ空間全体でのユーザの行動パターンを解析する研究。

3) 検索エンジンなどで入力された検索語（検索ログ）を用いて、人々のニーズやトレンドの抽出を行うことを目的とした研究。

また、ブログや日記に関してはmixi やGREE のようなSNS の登場により、ここ数年で爆発

的に普及している。ブログや日記には製品、レストラン、観光地、テレビ番組などの評判などが多く含まれているため、これらの情報から評判や流行を抽出すべくブログマイニングと呼ばれる研究が企業も含め盛んに行われている。

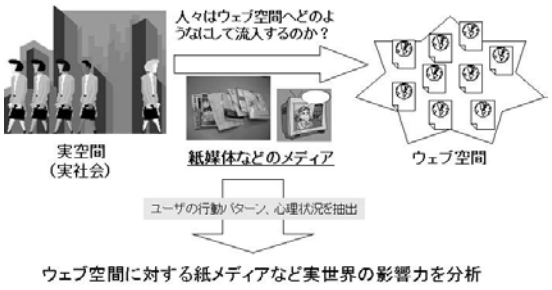

図１　研究の目的

２．研究の目的

本研究では図１に示すように、フリーペーパーやテレビなど、身近なメディアからの情報をもとにインターネットへアクセスを行うユーザの行動解析や分析を行い、その結果からウェブ空間と実世界を動く人々の行動パターンを明らかにし、ウェブ空間に対する紙メディアなど実世界の影響力を分析することを目的とする。

研究の全体構想としてはフリーペーパーと連動するホームページのアクセスログの解析を行い、紙メディアを利用してウェブ空間を訪れたユーザの行動解析を行う。

また、テレビに関連するメタデータ（番組表情報）やブログ記事、日記をウェブ空間から収集することで、テレビで放映されたドラマやスポーツ、生活情報などがユーザの行動にどのような影響を与えているかの分析を行う。これらの結果から得られた知見により実世界がウェブ空間に及ぼす影響力についての検討を行い、さらにはユーザの社会的な行動パターンの解明を行う。

３．研究の方法

まず、紙メディアと連動するホームページのアクセスログの解析、および、ウェブ空間からテレビ番組のメタ情報や関連するブログ・日記の収集を行う。次に、収集したデータの解析とブログ・日記に書かれた内容の検討を行う。さらに、これまでに得られた様々な知見から、紙メディアやテレビなどの実世界がウェブ空間に及ぼす影響力について分析・検討を行う。

(1) 平成２１年度の研究方法

テレビ番組情報とブログ･日記についてはサイトの調査と試験的な収集を行い、どのような情報を収集し分析を行うかを検討する。また、女性向けフリーマガジン「Well」の発行サイトについて紙メディアとそれに連動するホームページのアクセスログの解析を行う

Wellは偶数月の20日に発行され、毎号約30万部が繁華街での街頭、地下鉄の駅やコンビニ、千代田･中央･港区の約4000オフィス、飲食店･スクール･美容院などの協力店などで配布されている。このサイトのアクセスログについて以下の処理や検討を行う。

- アクセスログ中の個人情報の除去やマスキング処理
- アクセスログの時間的傾向の分析
- アクセスログ中に含まれる検索後の分析
- アクセスログ中に含まれるユーザ行動の分析

(2) 平成２２年度と平成２３年度の研究方法

テレビに関するブログ記事を収集し、分析を行いテレビ番組のブログへの影響を検討する。また、女性向けフリーマガジン「Well」の発行サイトについての解析結果を元に検索エンジンやポータルサイトの検索語の頻度情報との比較･検討を行う。検索語の頻度に基づく予測変換が検索サイトなどで行われており、サイト内検索でも有効であると考えられる。このため、サイト内検索における検索語の頻度を元に検索語の予測変換システムについて検討を行う。さらに、テレビ番組のウェブへの影響については、テレビ番組について記述のあるブログを判別し分類する検討を行う。また、テレビ番組の放送日とブログで話題にある数の時間的変化をとの関連を検討する。

４．研究成果

本研究課題の期間中に、テレビ番組のウェブへの影響についての調査･分析を行うため、テレビ番組について記述のあるブログを判別し分類する検討を行った。また、テレビ番組の放送日とブログで話題にある数の時間的変化をとの関連を検討した。

また、女性向けフリーマガジン「Well」の発行サイトについて紙メディアとそれに連動するホームページのアクセスログの解析結果から、ウェブ空間に対する紙メディアなど実世界の影響力を分析することができた。また、ユーザの行動分析をより詳細に行うために、サイト内検索における検索語の頻度を元にした検索語の予測変換システムに関する実験や、フリーマガジンサイト内における店舗推薦に関する実験を行った。ここでは、代表的な研究成果について述べる。

(1) テレビ番組に関するブログの分類と番組推定

インターネット上のブログを1)テレビに関するものか、2)どのジャンルに属するか、について自動分類を行い、さらに記述されている番組内容を特定することで、ブログの作者（ブロガー）へ対するテレビ番組の影響を調査した。

① ブログの収集

最初に Google ブログ検索で無作為にブロ

グを収集する。ここで、83 種類の平仮名を 1 文字ずつ検索語として設定し、1 日分の検索結果を収集した。次にブログの HTML ソース内から本文とコメントを抽出する。DIV タグを境としてソース内の文章を区切り、文章らしさとしての重みを付与する。重みの計算には、加算対象に句読点と改行タグの数を、減算対象にリンクの数と特定のキーワードを用いている。

② ブログの分類

Yahoo! 掲示板から投稿を収集し、テレビ関連のものとそれ以外を分けてデータベースへ学習させる。テレビ関連の投稿については、「特撮ヒーロー」「アニメ」「スポーツ」「ドラマ」「音楽番組」「時代劇」「子ども」「バラエティ」「ニュース」「コマーシャル」「全般」の 11 種類のジャンルへ分けて別のデータベースへ学習させる。ブログから抽出した本文・コメントを形態素に分割し、データベースに保存してある学習データを用いてナイーブベイズ分類器でテレビ関連の判定とジャンルの自動分類を行う。

③ ブログの番組特定

1 年分の番組表からタイトルを抽出して形態素に分割し、番組特定に使用するための索引を作成する。ブログを形態素に分割した文字列を検索語として索引から番組名を検索し、tf-idf 値の高いものを、そのブログと関連性の高い番組名として採用する。

④ 実験結果

ある 1 日分のブログ 5､613 件に対してテレビに関するものを検索し、得られたブログ 1､093 件に対してジャンルの分類を行った結果を図２に示す。テレビに関するものを検索した際の再現率はある程度高かった。また、番組特定では、番組名が略称で記述されていた場合でも検索することが可能であった。

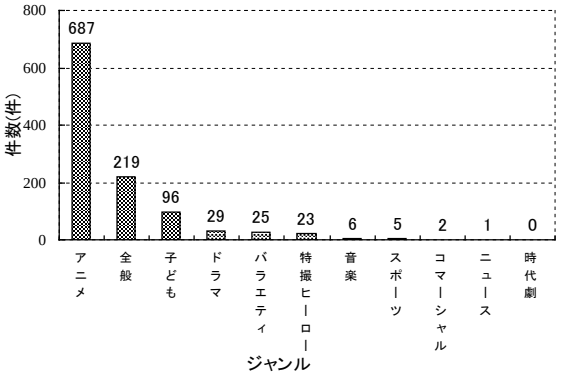

図２ テレビ関連のブログにおけるジャンルの件数

収集したブログのうちテレビに関するブログであると判定されたブログは 10 数%であった。判定の適合率・再現率については 70%から 80%程度であったが学習件数を増やすと向上する。ジャンル分類では娯楽番組に関するブログが多いことが分かった。番組放送日とブログ件数の関連については放送当日か翌日に増える番組が多いが、特定日に集中しない番組もあった。また、特集が行われた週などにアクセスが多い現象が見られる番組もあった。

(2) 女性向けフリーマガジンと連動するサイトにおけるユーザの行動分析

人々はある周期で睡眠や食事を取り、多くの人は仕事をしているため、Web 空間における行動は実世界と何らかの関連性を持っている。そこで、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のアクセスログを解析することで、実世界とWeb 空間における行動の関連性を見出すことを試みた。

① フリーマガジンとアクセスログデータの詳細

我々が解析を行ったサイト1と連動したフリーマガジン「Well」について説明を行う。この雑誌は偶数月の20日に発行され、毎号約30 万部が以下の方法により配布されている。

- 朝と夕方、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 店舗)

雑誌の対象者は銀座・丸の内・青山・六本木など山手線の内側で働く女性とし、紹介されている店舗は、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、リラクゼーションの8つのジャンルに分類される。

雑誌に掲載された店舗情報は連動するサイトに随時掲載される。このサイトではユーザ会員登録などは行っていないため、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から20-30 代の女性が大多数を占めていると考えている。今回の解析に用いたアクセスログは2004 年2月から2008 年12 月のものを対象としており3､Web クローラーなどのアクセスを除外するなどデータのクリーニング処理を行っている。各店舗ページにはWell コードと呼ばれるIDを割当てており、さらに、Well コードからジャンルを特定することが可能である。また、特集記事ページからもジャンルを特定することが可能なため、今回の解析ではジャンル毎にアクセス状況を調査した。

② 解析結果

アクセスログから得られる情報を元に「アクセス時間とジャンル」「アクセス曜日とジャンル」の関連性について調査を行った。軸はアクセス数を示しているが、実数についてはここでは非公開とする。

● アクセス時間とジャンルの関連性

アクセス時刻とアクセス数の変化について、各ジャンルにまとめたものを図３に示す。横軸はアクセスした時間帯を示し、縦軸はその時間帯のアクセス数をそのジャンル全体のアクセス数で割った値である。どのジャンルも朝から昼にかけてアクセス数が増加し、13 時のアクセス数が減少する傾向がみられる。その後、グルメ、レジャー、リラクゼーションは増加に転じ17 時台にピークとなる。ビューティー、ショッピング、クリニック、ヘアは夕方まで減少し、その後、夜になって増加に転じる。キャリアについては15 時台をピークに減少に転じる。

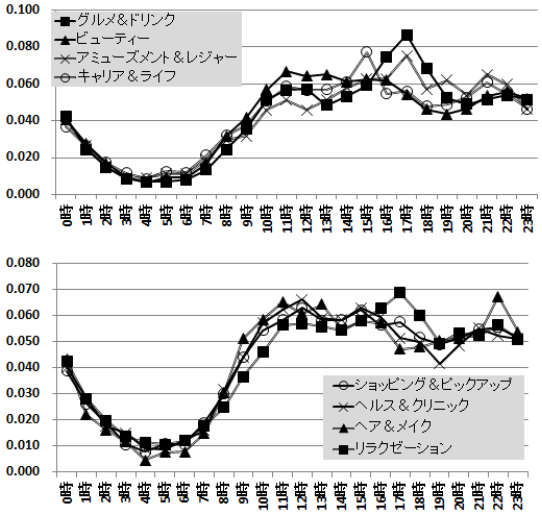

図３ アクセス時間とジャンルの傾向

● アクセス曜日とジャンルの関連性

アクセスした曜日とアクセス数の変化について、各ジャンルにまとめたものを図４ に示す。横軸はアクセスした曜日を示し、縦軸はその曜日のアクセス数をそのジャンル全体のアクセス数で割った値である。どのジャンルも土日のアクセス数が低いが、日曜日のレジャーと土曜日のヘアのアクセスは他のジャンルに比べ少し高いことがわかる。各ジャンルの特徴として、1) グルメは火曜日から金曜日のアクセスが多い。2) ビューティーは火曜日がピークであり、その後は減少傾向である。3) レジャー、ショッピング、クリニックは週初めのアクセスが多い。4) キャリアは水曜日がピークであり、その後は減少傾向である。5) ヘアは木曜日と金曜日のアクセスが多い。6) リラクゼーションは火曜日、水曜日、金曜日のアクセスが多いなど、ユーザの特徴的な行動を抽出することができた。

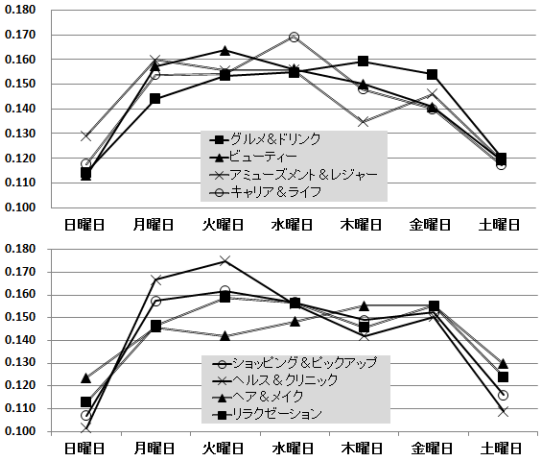

図４ アクセスした曜日とジャンルの傾向

(3) 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

前述のフリーマガジン発行サイトにおける店舗ページ推薦を行うために、アクセスログの解析を行い、似たような趣味を持つユーザを発見する手法の検討を行った。ユーザは手間をかけずに自分の好きな店舗が推薦されるため、サイトのアクセス数向上に貢献できると考えている。

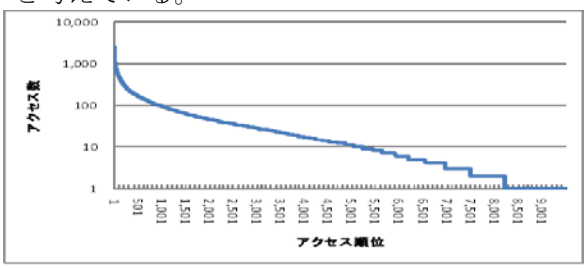

図５ 店舗ページのアクセス数の傾向

① 店舗ページのアクセス数の傾向

店舗ページのアクセス数についての調査を行い、アクセスが多い順に店舗を並べた結果を図５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わかる。

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み合わせの調査を行った。ユーザが閲覧した店舗ページの組合せについて調査を行った。組合せ頻度の傾向を図６に示す。図の横軸は組み合わせ頻度が多い順に並べた時の順位であり、縦軸は頻度を表している。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わかる。

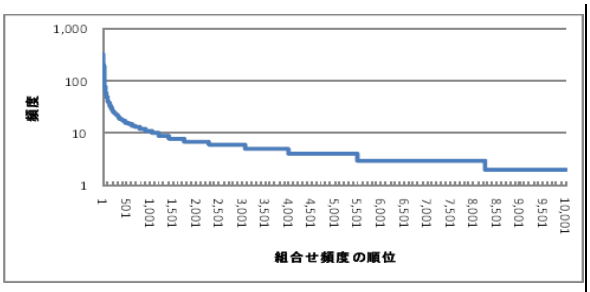

図６　同一セッション内で閲覧された共起する店舗ページのアクセス傾向

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た。その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を閲覧することが多いことがわかった。また、ジャンル「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。

(4) サイト内検索における検索ワードの予測システムの構築

あるサイトにおいてアクセスログの解析を行い、検索ワードの予測変換を行うシステムを作成する。その際、検索ワードは時間により頻度の偏りがあると考えられるため、曜日・日時ごとに表示されるワードを変化させることで、よりユーザにとっての利便性を高めることを目指した。

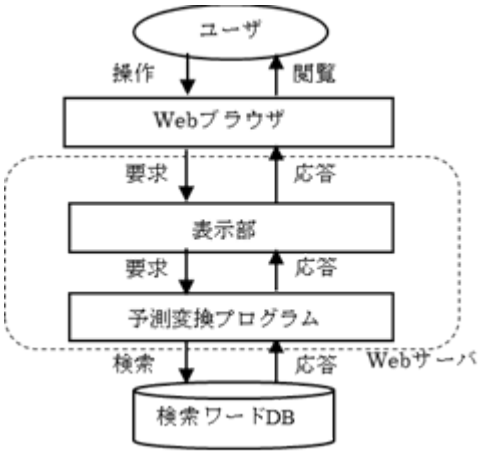

図７　システム概要

我々が構築した予測変換システムでは、前述のフリーマガジンサイトのアクセスログより時間帯ごとの検索ワードの頻度を調べ、時間帯ごとに高い順に検索ワードを表示させる。時間帯の他に曜日や季節などが選択が可能である。予測変換システムの前段階として、アクセスログから検索ワードと時刻を抽出してデータベースに格納しておく。

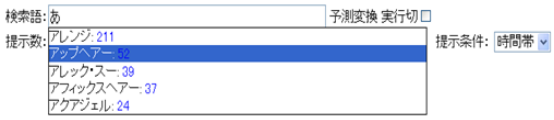

図８　「あ」と入力した時の予測変換

本研究のシステムを用いることで、ユーザの入力の手間やミスを減らすことができると考えられる。

５．主な発表論文等

〔雑誌論文〕（計１件）

① 大塚真吾，宮崎収兄：女性向けフリーマガジンと連動するサイトにおけるユーザの行動分析，日本知能情報ファジィ学会誌「知能と情報」，査読有，Vol.3，2012，採録決定

〔学会発表〕（計３件）

① 大塚 真吾，高橋 修太郎，宮崎 収兄：女性向けフリーマガジンと連動する Web サイトのユーザ行動解析，第 17 回 Web インテリジェンスとインタラクション研究会，Vol.WI2-2010，No.1-31，pp.59-60，2010 年 3 月 15 日，大阪大学中之島センター

② 高橋 修太郎，大塚 真吾，宮崎 収兄：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、第 2 回データ工学と情報マネジメントに関するフォーラム(DEIM2010)，I-1，2010 年 2 月 28 日，淡路夢舞台国際会議場

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，第 1 回データ工学と情報マネジメントに関するフォーラム(DEIM2009)，B8-4，2009 年 3 月 9 日，ヤマハリゾートつま恋

６。研究組織

(1)研究代表者
宮崎　収兄（MIYAZAKI NOBUYOSHI）
千葉工業大学・情報科学部・教授
研究者番号：20265466

(2)研究分担者
大塚　真吾（OTSUKA SHINGO）
神奈川工科大学・情報学部・准教授
研究者番号：70509736

高久　雅生（TAKAKU MASAO）
独立行政法人 物質・材料研究機構・科学情報室・主任エンジニア
研究者番号：00399271