社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－DBS－134　(I)　(4)
2004／7／13

アクセス履歴とディレクトリ型検索システムを用いた利用者集合の興味動向分析

平野 真太郎† 吉岡 由智† 成 凱‡ 岩井原 瑞穂†

†京都大学情報学研究科 〒606-8501 京都市左京区吉田本町
‡九州産業大学情報科学部 〒813-8503 福岡市東区松香台 2-3-1

E-mail: †{shin, yoshitomo, iwaihara}@db.soc.i.kyoto-u.ac.jp, ‡chengk@is.kyusan.ac.jp

あらまし ウェブの大きな特徴として，ホットスポットとよばれる少数のウェブサイトに利用が大きく偏っていることが分かっている．膨大なウェブの効率的な活用のために我々は利用者の興味を反映したホットトピックを検出するトピックセンサー[1]を提案してきた．しかし，トピックごとの利用頻度を計算する際トピックの階層構造を考慮しておらず，ホットトピックの検出手法は単純である．本稿ではトピックセンサーの拡張としてHHH(Hierarchical Heavy Hitters) アルゴリズム[2]を用いてトピックの階層構造をより効果的に扱う方法を提案する．Yahoo!カテゴリからトピックの階層構造を抽出し，アクセス履歴における利用状況を考慮することによって時間軸を考慮したホットトピックの検出を行った．トピックの特徴，例えば朝によく利用されるトピック，夜間によく利用されるトピックなどの時間帯による利用状況が知ることができれば，インターネット広告においてより効果的な広告,高度な課金システムの作成が可能になると考えられる.実験では時間帯によるトピックの利用状況を解析し，時間帯によってトピックが 3 つのタイプに分かれることを確かめた.

キーワード トピック，アクセス履歴，ディレクトリ型検索エンジン，インターネット広告，HHH(Hierarchical Heavy Hitters)

Analyzing User Interest by using Access Log and Directory of the Web

Shintaro HIRANO† Yoshitomo YOSHIOKA† Kai CHENG‡ and Mizuho IWAIHARA‡

†Graduate School of Informatics, Kyoto University, Yoshidahonmachi, Sakyo-ku, Kyoto, 606-8501
‡Faculty of Information Science, Kyusan University Matukadai 2-3-1, Higashi-ku, Fukuoka, 813-8503

E-mail: †{shin, yoshitomo, iwaihara}@db.soc.i.kyoto-u.ac.jp ‡chengk@is.kyusan.ac.jp

Abstract A salient feature of the web is its biased usage where a few hot spot sites account for most accesses. To detect the hot topics that reflect users' interest, we have developed a system, called topic sensor [1]. However, in that work, we did not take into account the hierarchical structure while counting the access frequencies of each topic. In this paper, we extend that work by allowing topic hierarchy and detect hot topics by Hierarchical Heavy Hitters (HHH) detection algorithms of [2]. We adopt topics Yahoo! directory, and determine their "hotness" by using access logs from shared proxy servers. We demonstrated that a special feature of the usage of topics, for example one topic is often accessed in the morning, the other is often accessed in the night, can improve current Internet Advertisement. We report the result of analysis by hours and ensured that Topics are enabled to divide into 3 types.

Keyword Topic, Access Log, Directory of the Web, Internet Advertisement, HHH(Hierarchical Heavy Hitters)

1. はじめに

現在のインターネットの世界では，ウェブデータは膨大な量になっている．そしてこれからも増加の一途をたどっていくと考えられるが，このような環境下では利用者が自分の必要とするデータだけを取得するのは困難である．従って，利用者の WWW 上での活動を支援するために，利用者から必要とされているデータを検出することが重要になる．我々はこの状況を解決するために，ウェブデータの内容やその利用状況を考慮して利用者の興味が集中しているデータを検出するトピックセンサー[1]を提案してきた.

[1]ではトピックの「詳細度」というトピック階層に近い概念について言及したが，トピックごとの利用頻度を計算する際，トピックの階層構造を考慮しておらず,ホットトピックの検出手法は単純である.しかし，実世界においてはトピックを階層構造としてその利用頻度を計算する必要がある.例えば，「俳優」というトピックの下に「男性俳優」と「女性俳優」などのサブトピックが考えられる.「男性俳優」か「女性俳優」のいずれがホットと判断されればよいが，両方ともホットの基準に至らない場合は親トピックである「俳優」が「ホット」と判断されうのは自然である.

本稿ではトピックセンサーの拡張としてトピックの階層構造を利用したホットトピック検出方法を提案する.

本研究の特徴は以下の 3 つである．

1. 階層構造を考慮したトピック利用頻度の計算
2. 情報量(トピックごとのウェブサイト数)と利用頻度を両方考慮したホットトピック分析
3. より詳細な時間帯別のトピック分析

我々はトピックの階層構造を抽出するために Yahoo!カテゴリ(http://www.yahoo.co.jp)を利用した．Yahoo!は代表的なディレクトリ型検索システムであり，トピックごとにカテゴリを持つ．Yahoo!カテゴリからトピックの階層構造をとりだし，トピックに含まれるサイト数を情報量とした．そして一般 ISP の会員のアクセス履歴を用いて各トピックの情報量に対応した実際の利用頻度を調べ，トピックを抽出し利用者の興味の動向を分析した．

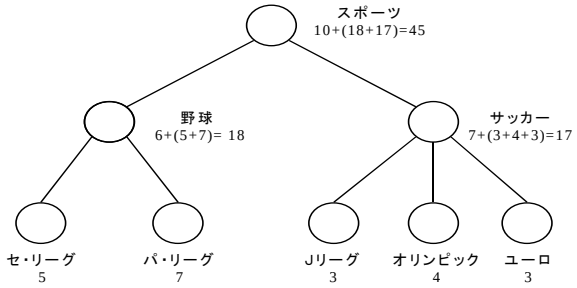
Yahoo! はポータルサイトとして利用されることが多く，ニュースサイトよりも利用が多いためより利用者の興味を反映したホットトピック抽出が期待できる．利用頻度の計算ではトピックへの単純なアクセス回数とせず，計算時間が短くメモリ使用量が小さい HHH(Hierarchical Heavy Hitters)アルゴリズム[2]を用いて階層構造を持つトピックに対応したアクセス集計を行った．

またトピックの利用の特徴，例えば朝によく利用されるトピック，夜間によく利用されるなどの時間帯による利用状況，そしてトピック同士の関連性などが知ることができれば，インターネット広告において広告効果に見合った料金システム，ならびにより効果的な広告作成が可能になると考えられる．例えば，表示時間に応じて料金が決める方式の場合，昼によく見られるトピックであれば，朝のうちに新しい広告を準備することによって，より効率的に多くの利用者に訴えることができるようになる．昼の時間帯にだけ広告を出すことによって宣伝コストを下げるといったことも可能になる．また Google(http://www.google.co.jp)などの検索エンジンにおいて特定のキーワードの検索結果にスポンサーサイトという形で広告を出す方法がある．例えば「C 言語」というキーワードの検索結果に対し，C 言語の e-learning サイトが現れるといった具合である．キーワードとその関連するトピック(例．C 言語と学習)を時間軸によって把握できれば，それを考慮した広告をだすことによってより効率的な効果が期待できると考えられる．実験では時間帯によるトピックの利用状況について分析した．

2. 関連研究

膨大なウェブデータを効率的に扱い利用者にわかりやすい形で提供することを目的とする研究が多数存在する．トピックセンサーのほかにも TDT(Topic Detection and Tracking)[4]はマスコミのニュースサイトを利用した研究があり，発信するニュースデータに着目し，そのデータを効率的に扱い利用者にわかりやすい形で提供することを目的とする研究である．我々は Yahoo!トピックというあらかじめトピックが分類され，かつ利用度の高い情報を利用することでトピックを分類するコストを省いている．

インターネット広告の効果測定の手法としてユーザーセントリックと呼ばれるものがある．これは調査会社が，インターネット利用者の性別などの属性を反映させた調査用パネルを抽出し，測定用プログラムのインストールを依頼して，そのアクセス履歴を解析するというものである．複数のウェブサイトやインターネット広告の効果を同一基準で比較できる長所を持つものである．このアクセス履歴を利用し利用者の大域的な行動を把握することを目的とした研究[3]がある．利用者が区別できる貴重なアクセス履歴を用いており，類似するウェブサイトをまとめる技術であるウェブコミュニティとアクセス履歴に残る検索結果を用いたウェブログ解析システムを提案している．



	野球	サッカー
情報量	3	4
利用頻度	18	17

図 1．分析手法の基本的概念

3. HHH を用いたホットトピック分析

3.1. 概念

提案する手法の特徴となる 2 つの概念を説明する．

- 情報量
  トピックに含まれるウェブサイトの数．人気のあるトピックほどサイト数が多い．
- 利用頻度
  実際に利用者によって利用された頻度．よく利用されたトピックほど重要である．

この情報量を考慮した利用頻度によってホットトピックを検出する．ここではトピックの利用頻度を計算する際に用いる手法について図 1 を利用して説明する．

まず Yahoo!カテゴリからトピックの階層構造を抽出する．図 1 の木構造がトピックの階層構造を表して

いる．トピックの情報量は子孫トピックの数であり，図 1 の野球トピックの情報量は 3 である．次に利用頻度をウェブサイトへのアクセス回数を用いて計算する．図 1 の野球はそのサイト自体のアクセス回数(6 回)と子であるセ・リーグ(5 回)とパ・リーグ(7 回)のアクセス回数(利用頻度)を加えたものとなり 18 となる．つまり各トピックの利用頻度は子孫トピックの利用頻度を考慮したものになる．情報量が多く，かつそのアクセス回数が多いトピックの利用頻度が大きくなる．

しかし，図１のように子孫トピックの利用頻度の単純加算では上位のトピックの利用頻度が大きくなるばかりで有用なホットトピックは検出できない．そこでトピックの利用状況の計算を階層構造の特徴を生かして計算できる HHH(Hierarchical Heavy Hitters)[2]によって行う．

3.2. 頻出階層構造検出アルゴリズム HHH

HHH はネットワーク管理などの通信が大量発生するアプリケーションにおいて頻出するアイテムのことである．本研究では利用者によるウェブサイトアクセスをアイテムとして捉えトピックの利用状況を計算した．以下に階層型トピック分析に適応させた HHH について述べる．

階層型ドメイン D による多重集合 S を準備する．次のように HHH を定義する．各トピック p の利用頻度を F(p)，$N=\sum pF(p)$とする．$\phi \in (0,1)$は閾値である．

HHH0：頻出している S の要素 e　　$f(e) > \phi N$

HHH_i：頻出している階層 i の接頭 p　　$F(p) > \phi N$

$F(p)=\sum f(e): e \in elements(\{p\}) \wedge e \notin elements(\bigcup_{j<i} HHH_i)$

ここで，elements(T)は子孫関数と呼び，接頭集合 T の全ての子孫要素の集合を求める関数．F(p)は子孫の利用頻度も加えたもので，かつ子孫の中の HHH を除いたものとなる．

実行例は次の通り．トピック“コミックとアニメーション”の利用頻度を求めるため，“コミックとアニメーション”の“コミック”が頻出と判断された場合には，その分の利用頻度を“コミックとアニメーション”から引く必要がある．理由は「その分のターゲットは，“コミック”であり，一般的な“コミックとアニメーション”ではない」からである．

次に階層構造を持つトピックから HHH を検出するためのアルゴリズムを以下に示す．$b_current = \lfloor \varepsilon N \rfloor$で N はその時点での総アクセス数，初期値は 0．$\varepsilon \in (0,1)$は誤差範囲．以下のアルゴリズムは階層構造が木構造をしており，挿入される要素(節)が親を覚えていることが条件である．

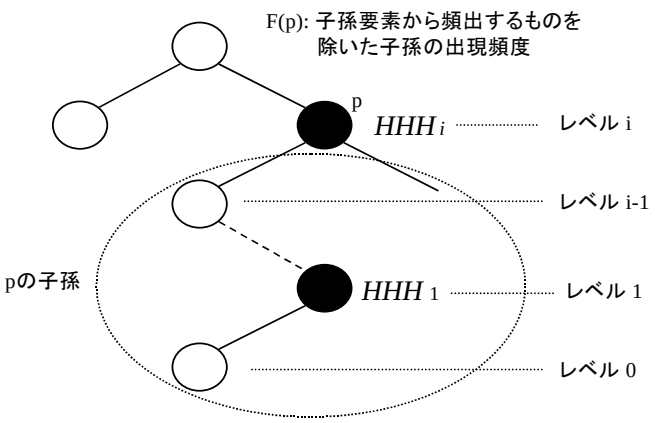


図 2．HHH の仕組み

HHH 検出アルゴリズム

1. $w=\lceil 1/\varepsilon \rceil$個の要素を順に挿入する
   (1) 新しい要素の時
      ① 要素から根までの trie 構造を作る(要素は親を覚えている)
      ② Δ =b_current (N はこの時点でのΣ pf)
      ③ 要素から根までの節の値 f に 1 を加算
   (2) trie 構造に要素である時
      ① 要素から根までの節の値 f に 1 を加算
2. 葉の要素に限り f + Δ < b_current であればその要素を削除する
3. 1 に戻る

出力：$f+\varepsilon N > \phi N$を満たす要素がある時

(1) HHH として F を出力する
(2) 出力した要素の親の F から f を除く(枝を切る)
(3) 親が $f+\varepsilon N > \phi N$を満たす時(1)を行う

HHH 検出アルゴリズムは要素の数が膨大で，処理時間や，再読み込み不可能などの制限で，正確な計算に必要なメモリ量は足りていない場合を想定した近似なものである．

そして時系列とそれに関係する人気トピックを調べるために朝，昼，夜，深夜の時間帯に分けて利用頻度を集計した．トピックの時間帯による利用頻度を調べるために，以下のように 6 時間刻みの時間帯を定義した．

1．朝　5:00～10：59
2．昼　11：00～16：59
3．夜　17：00～22：59
4．深夜　23：00～4：59

このように時間帯別の利用状況を調べることによって昼間に人気のあるトピック，深夜に人気のあるトピッ

クなどを知ることができればバナーなどによるインターネット広告において大きな助けとなると考えられる.

4. 実験及び考察

本研究では Yahoo!というディレクトリ型検索エンジンのトピックの階層構造(Yahoo!トピック)に注目し,それに対する利用頻度を時間帯別に計算し利用者の興味の動向を分析した. また, 1 週間単位の利用状況の推移を 4 週に渡って調べその人気トピックの推移も観察した. 解析には Perl を利用し, トピック構造抽出とテキストタイプのプロキシログ処理行った.

4.1. Yahoo!カテゴリからのトピック抽出

Yahoo!カテゴリは“エンターテイメント”や“趣味やスポーツ”といった主要なトピックによって構成されている. 興味のあるコンテンツ(リンク)を辿っていくことによってより詳細なトピック情報にアクセスすることができる. Yahoo!は手動によってサイトがトピック分類され登録されているため, その分類は信頼できると考えられ有用な情報源である. トピック情報を使うことによってトピックの抽出, 分類のコストを省くことが出来る.

Yahoo!カテゴリの構造について説明する. 図 1 に示すように Yahoo!トピックでは上部にトピックの説明がある.「エンターテイメント＞芸能人,タレント＞アイドル」と言った具合である. ページに含まれる主要な情報として次の 2 つがある.数字は 04 年 6 月 4 日時点のものである.

- Yahoo!カテゴリ
  子のトピックへのリンク.「エンターテイメント＞芸能人, タレント＞アイドル＞イベント」という具合である. エンターテイメントに関するサイトは 11,027 種ある.
- Yahoo!登録サイト
  そのトピックに関するサイトで”yahoo”をアドレスに含まないもの. エンターテイメントに関するもので 38,128 種あった.

Yahoo!カテゴリにおいて登録サイトは一つのカテゴリサイトにのみ表れ複数表れることはない. またリンク先は他のトピックへまたがることも多い.例えば「エンターテイメント→地域情報」といった具合である. トピック毎の単純な木構造ではなく複雑な構造をしている. 構造抽出において多重ハッシュを利用した.

4.2. アクセス履歴

本研究では京都市の ASTEM(京都高度技術研究所)の運営する ISP, Kyoto I-net のプロキシログを利用して実験を行った. 抽出したトピックの利用状況の計算にはプロキシログを用いる. プロキシログはウェブ上での利用者の活動(アクセスした URL)を時間順に保持

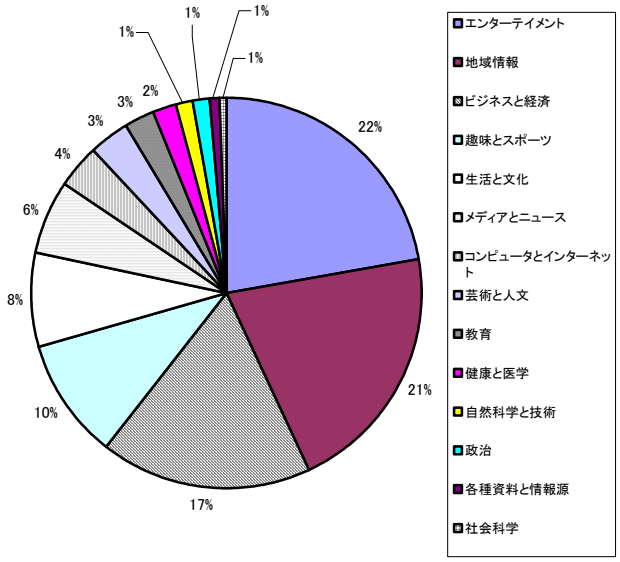


図 3. 主要トピック別利用状況

したものである.KyotoI-net の会員数は 2 万人以上で, 解析に利用したデータは 03/11/01 から 03/11/30 の 1 ヶ月分である. 28 のプロキシサーバが稼動しており, 各々が独立にプロキシログを記録している. 実験に利用したレコード総数は 56,761,653 にも及ぶ.

プロキシログに現れるファイルタイプが text/html であるものに限った場合, そのレコード数は 11,601,974 であり全体の 20.4%を占める. 実験ではこの tex/html タイプのレコードを利用した.

Yahoo!に関連するサイト(“yahoo”を URL に含むもの)は 1,285,504 あり, 全体の 11.1%にもなる. このことは Yahoo! が利用者によって如何によく利用されているかがわかる. そのうちの主要コンテンツへの利用状況を表 1 に示す.

Yahoo!	1,285,504
Yahoo!オークション	459,154(35.7%)
Yahoo!検索	188,460(14.7%)
Yahoo!掲示板	101,588(7.9%)
Yahoo!トピック	19,689(1.5%)
Yahoo!ホームページ	81,194(6.3%)

表 1. Yahoo!の利用状況

Yahoo!オークションや Yahoo!検索, Yahoo!掲示板のようにインタラクティブ性のあるコンテンツの利用が多く全体の 60%近くにもなることが分かった. その多くは検索が利用されていることが分かった. 次に Yahoo!トピックの中の 14 の主要トピックの利用状況を図 3, アクセス回数を表 2 に示す. 図 3 と表 2 からエンターテイメントを初めとする上位のトピックにアクセスが集中していることが分かる. これらの結果より, 実験では最も利用の多かったエンターテイメントトピックを対象とした.

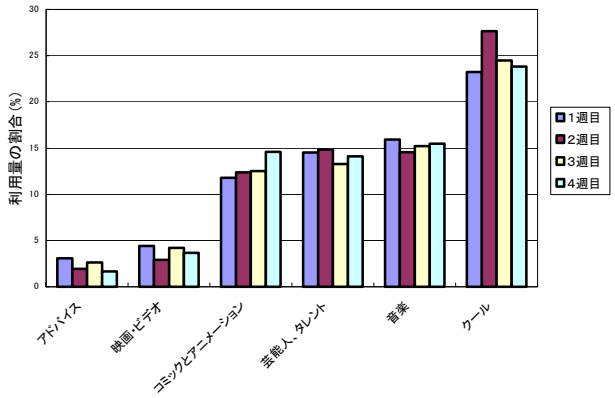


図 4. エンターテイメントトピックの利用頻度

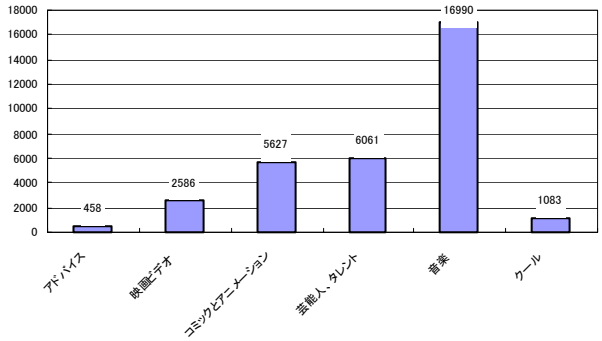


図 5. 主要トピックの情報量

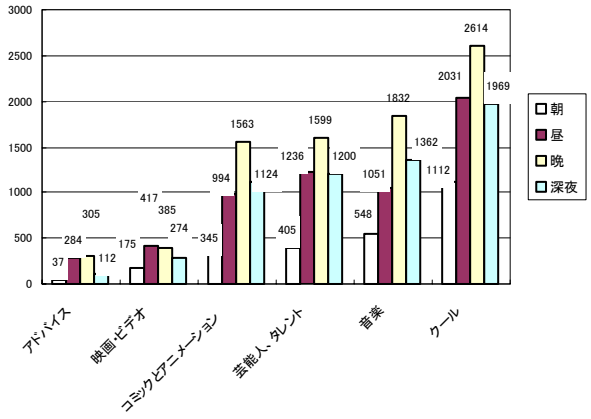


図 6. 主要トピックの時間帯毎の利用頻度

順位	トピック	アクセス回数
1	エンターテイメント	4371
2	地域情報	4107
3	ビジネスと経済	3418
4	趣味とスポーツ	2012
5	生活と文化	1497

表 2. 主要トピックのアクセス回数

サイト情報は平成 16 年 6 月 4 日に取得したものである．データを木構造に整形して利用した．

利用したアクセス履歴に現れたエンターテイメントの子となる Yahoo!カテゴリと Yahoo!登録サイトはそれぞれ 1122 種，4613 種であった．実験においてはこの 5247 種のトピックへの利用頻度を解析した．なおプロキシログに現れる IP アドレスは利用者を特定できないように暗号化されたものを用いることにより，Kyoto I-net の会員のプライバシーには十分注意した．

4.3. 評価および考察

エンターテイメントトピックの 4 週に渡る利用状況を図 4 にしめす．横軸は子であるとトピック，縦軸はその週の全体における各トピックの利用頻度の割合である．特にジャンルを問わず「クール」な情報が集められているクールトピックがよく利用されていることが分かる．これはクールトピックに含まれる登録サイトがよく利用されていたためであった．11 月の 4 週間の利用の移り変わりにおいては大きな変化がおきていないことが分かる．これはエンターテイメントの利用は時期に影響されにくいためと推測できる．今後対象をスポーツや経済といった時期に影響の受けやすそうなトピックで解析を行う必要がある．

図 5 はエンターテイメントトピックの子の中で主要なトピックの情報量(登録サイト数)を表している．音楽トピックが非常に多く，クールトピックに含まれるサイト数は少ないことが分かる．音楽トピックのサイト数が多いのは各歌手のトピックが存在するためである．一方，図 5 は図 4 であげた各トピック別の時間帯毎の利用頻度を示している．クールトピックへのアクセス数が音楽トピックへの利用頻度を超えていることが分かる． これらのことは，トピックのサイト数が多いことが人気トピックとは限らないということを示している．トピックの人気を調べる上でトピックの情報量だけでなくはなく，利用頻度を調べることの有効性を示している．

4 つの時間帯において最もアクセス頻度が大きかった時間帯は夜間で，全 5,247 のトピック中 5119 のトピックがそうであった．これは夜間にインターネットを利用する利用者が多いことに関係していると考えられる．しかし，128 のサイトは昼間もしくは深夜といった時間帯にもっともよく利用されていることが分かった．昼間利用型，夜間利用型，深夜利用型の例をそれぞれ図 7，図 8，図 9 に示す．横軸は時間帯，縦軸は全体に対する各時間帯の利用頻度の割合である．

図 10 は SMAP とモーニング娘のメンバーサイトの利用状況を示している．SMAP は昼間利用型であり，モーニング娘は深夜利用型である．同じ芸能人，タレントトピックであっても利用のされ方が違うことが分かる．トピックによって利用頻度がピークとなる時間帯が異なることは，トピックに興味を持つ利用者のウェブ利用スタイルの表れと捉えることができる．

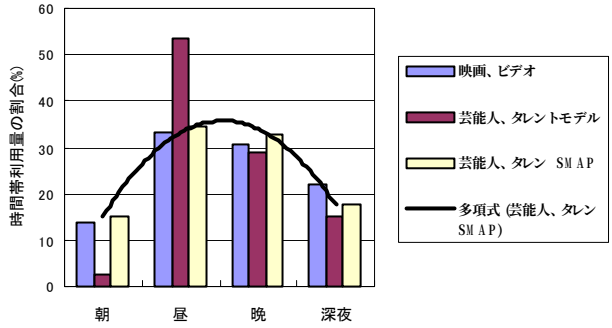


図 7. 昼間利用型の例

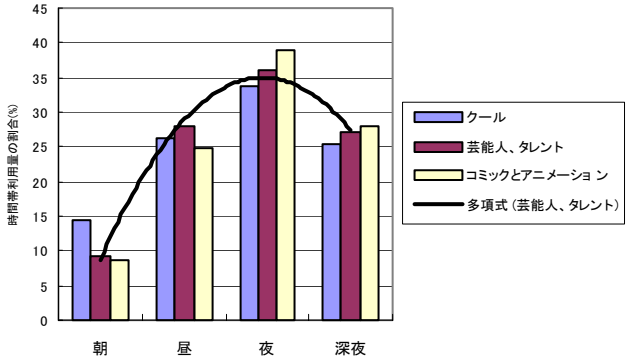


図 8. 夜間利用型の例

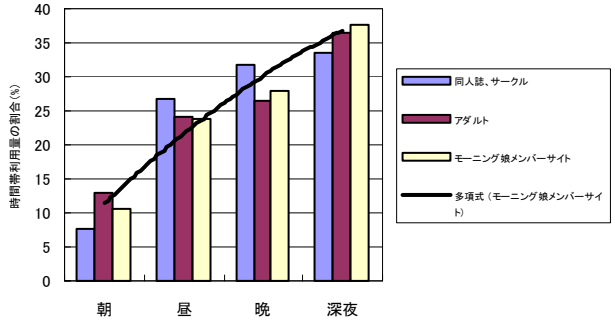


図 9. 深夜利用型の例

これはトピック同士の相関性，つまり利用者の興味の観点から見てどのトピックとトピックの関連があるのかを調べるうえで有益な情報となる．このような時間帯による人気トピックの情報は利用者の行動パターン把握と共にインターネット広告におけるより高度な課金システムの構築に応用すること可能になると考えられる．そのためには今後より信頼性のある実験結果を示す必要がある．またほとんどのトピックの利用頻度の最大時が夜に集中していたため，さらに時間帯の単位を小さくすることにより利用状況をより詳しく調べることができると考えられる．最後に上記の図で利用したトピックの時間帯別の利用頻度を表 3 にしめす．

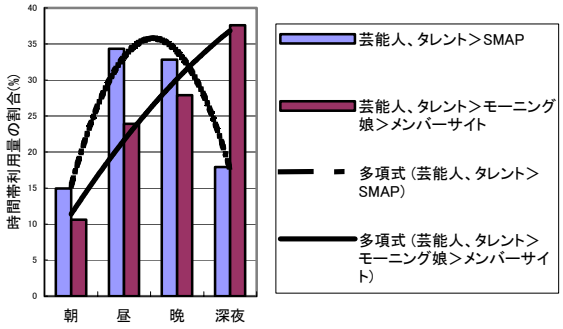


図 10. SMAP とモーニング娘メンバーサイトの利用状況の比較

トピック	朝	昼	夜	深夜	合計
映画，ビデオ	175	*417*	385	274	1251
芸能人，タレント＞モデル	3	*67*	*36*	19	125
芸能人，タレント＞SMAP	20	*46*	44	24	134
コミック，アニメ＞同人誌, サークル	30	104	124	*131*	389
その他＞アダルト	770	1444	1601	*2201*	6016
芸能人，タレント＞モーニング娘＞メンバーサイト	24	54	63	*85*	226

表 3. 各トピックの時間帯別利用頻度

5. まとめ

本稿では階層構造を持つトピックを扱い，その利用頻度を分析しホットトピックの検出および分析を行った．朝，昼，夜，深夜という時間帯とトピックの利用状況関係を調べることで，時間軸を考慮した人気トピックの抽出およびインターネット広告への利用の可能性について示した．今後はエンターテイメント以外のトピックも対象に含めた実験および，アクセス履歴の利用者を IP によって区別したより詳細な解析を行う予定である．

文　献

[1] 吉岡由智，平野真太郎，成凱，上林彌彦“ニュースサイトと履歴ウェブによるトピックセンサー，”DBSJ Letters vol.3 No.1

[2] G. Cormode, F. Korn, S. Muthukrishnan nad D. Srivastava, “Finfing Hierarchical Heavy Hitters in Data Streams,” 29th International Conference on Very Large Data Bases (VLDB),2003

[3] 大塚真吾，豊田正史，喜連川優 “Web 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，” 情報処理学会研究報告，2003-DBS-131(Ⅱ),pp101—108,2003

[4] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang: “Topic Detection and Tracking Pilot Study Final Report,” Proceedings of the Broadcast News Transcription and Understanding Workshop1998