階層型カテゴリを用いたウェブサイトのアクセス履歴の時系列相関性解析

平野 真太郎† 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, iwaihara}@db.soc.i.kyoto-u.ac.jp, ‡chengk@is.kyusan.ac.jp

あらまし インターネットにおいてどのようなコンテンツが人気があり，またどのコンテンツ同士に関連があるかといった情報はインターネット広告会社をはじめとする非常に多くの会社が欲している情報である．しかし何十億といわれる膨大な量のウェブサイトを解析することは難しい．我々はウェブのインデックスといえるディレクトリ型検索システムの階層カテゴリとその利用状況を用いた効率的な利用者の興味動向分析を行う．ウェブサイトの時系列の利用頻度パターンの相関性を，類似検索や時間帯別の利用型への分類などにより階層構造を考慮して調べる．カテゴリの利用パターンの特徴，例えば朝によく利用されるカテゴリ，夜間によく利用されるなどの時間帯による利用状況，そしてカテゴリ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，昼によく見られるカテゴリであれば，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． 本稿では時間帯別の利用型分析とウェブサイトの利用頻度パターンの類似判断手法を用いたウェブサイトの相関性解析について述べる．

キーワード 階層型カテゴリ，類似判定，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

Time-series Similarity Analysis for Web Access Log Pattern by Using Hierarchical Category of Web

Shintaro HIRANO† 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, iwaihara}@db.soc.i.kyoto-u.ac.jp ‡chengk@is.kyusan.ac.jp

Abstract Information about user interest is very useful for many applications, especially Internet Advertisement. But it is difficult to analyze billons of web sites. To know user interest on web, we analyze directory of the web and its access log. Time-series similarity search in similar usage patterns considering hierarchy of web category help analyzing it. We demonstrated that a special feature of the usage of categories, 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 categories are enabled to divide into 4 types and propose

Keyword Hierarchical Category, Similarity Search, Access Log, Directory of the Web, Internet Advertisement

1. はじめに

何十億といわれる膨大な量のウェブサイトを解析することは難しい．Yahoo!カテゴリ[11]や DMOZ[8]といったディレクトリ型検索システムはインターネットの世界のインデックスを構成しているといえる．我々はディレクトリ型検索システムのインデックス構造，カテゴリの階層構造の利用状況分析による利用者の興味動向分析を行う．利用状況分析には一般のインターネット・サービス・プロバイダー(ISP)のアクセス履歴であるプロキシログを利用した．プロキシログは会員利用者のウェブ上の活動を記録したものであり貴重な情報源である．本研究の特徴は以下の３つである．

1. カテゴリの階層構造を考慮したカテゴリ利用頻度の計算
2. カテゴリの利用パターンによる類似カテゴリ分析
3. 一般 ISP の実データによる分析

我々はカテゴリの階層構造を抽出するために Yahoo!カテゴリを利用した．Yahoo!は代表的なディレクトリ型検索システムであり，カテゴリごとにカテゴリを持つ．Yahoo!カテゴリからカテゴリの階層構造を

とりだし，カテゴリに含まれるサイト数をサイト量とした．そして一般 ISP の会員のアクセス履歴を用いて各カテゴリのサイト量を考慮した実際の利用頻度を調べ，カテゴリを抽出し ISP の会員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，他のポータルサイトやニュースサイトよりもアクセスが多いため，より利用者の興味を反映した高頻度カテゴリ抽出が期待できる．

カテゴリの利用の特徴や共起関係を詳しく調べるために様々な角度から類似するカテゴリを調べる．とくにカテゴリの利用頻度パターンから得たフーリエ係数を利用した類似カテゴリ分析を行う．類似検索を利用し同じ利用パターンを持つカテゴリを探す．

またカテゴリの利用の特徴，例えば朝によく利用されるカテゴリ，夜間によく利用されるなどの時間帯による利用状況，そしてカテゴリ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，昼によく見られるカテゴリであれば，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[9]などの検索エンジンにおいて特定のクエリー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」というクエリーの検索結果に対し，C 言語の e-learning サイトが現れるといった具合である．クエリーとその関連するカテゴリ(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．

以下 2 章では関連研究，3 章ではカテゴリの時系列の利用頻度の計算方法について，4 章では時間軸やカテゴリの階層構造を考慮した相関性分析について述べ，5 章ではウェブサイトの利用頻度パターンの類似性を判定する方法について述べる．6 章では実験データと実験について説明する．実験では時間帯の利用の特徴についての調査と共起カテゴリの発見を行い提案手法の評価を行った．

2. 関連研究

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[12]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．

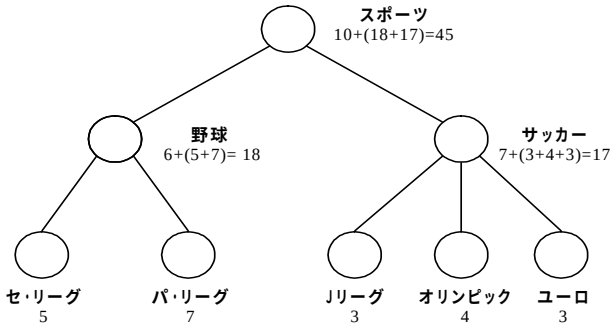
MSN[10]の検索エンジンの大量の利用結果を用いて，利用されたクエリー間の類似性を調べる研究がある[4]．各クエリーの時系列の利用データからフーリエ係数を求め，その係数を利用してクエリー間の類似性をユークリッド距離によって決定している．クエリーの利用周期性の発見およびバーストの発見の手法を提案している．バーストの例としてエルビス・プレスリーの命日付近でクエリー”elvis”がよく検索に利用されるといったことが分るという具合である．精度の良い類似クエリー検索を行うためにインデクシングの方法と効率の良いインデックス構造について複数手法提案し検証している．インデクシングに重点を置いた研究である．

3. カテゴリ利用頻度

分析の基本データとなるカテゴリの時系列利用頻度の計算方法において特徴となる 2 つの概念は次の通り．

- サイト量
  カテゴリに含まれるウェブサイトの数．人気のあるカテゴリ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．よく利用されたカテゴリほど重要である．

このサイト量を考慮した利用頻度を求める．ここではカテゴリ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．



	野球	サッカー
サイト量	3	4
利用頻度	18	17

図 1. カテゴリ利用頻度例

まず Yahoo!カテゴリからカテゴリの階層構造を抽出する．図 1 の木構造がカテゴリの階層構造を表している.カテゴリのサイト量は子孫カテゴリの数であり，図 1 の野球カテゴリのサイト量は 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．つまり各カテゴリの利用頻度は子孫カテゴリの利用頻度を考慮したものになる.このデータを分析単位(１時間，１日，１週間，１月)に時間推移と共に集計することによって時系列の利用頻度が計算できる．

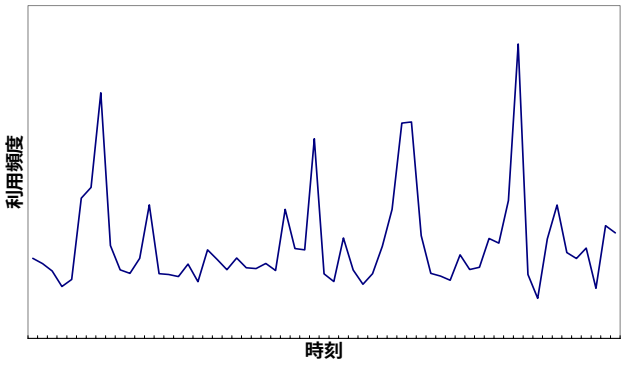


図 2. 時系列利用頻度パターンの例

4. カテゴリのアクセス相関性解析

カテゴリの階層構造を利用した分析を行う．クエリーは各サイトの時系列データから求めたフーリエ級数とし各対象カテゴリとの類似性を調べる．分析結果によって手動によって分類されたカテゴリ階層構造の見直しを行うことができる．利用者の違いやサーバの最適な配置方法などを見つけることも可能になる．

4.1. 時間帯の利用型による類似カテゴリ解析

時間帯別の利用状況を調べることによって，例えば朝によく利用されるカテゴリ，夜間によく利用されるなどの時間帯による利用状況，そしてカテゴリ間の共起関係などを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，バナー広告で一般的に利用されている表示時間に応じた掲載料金方式においては，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．

4.2. 兄弟カテゴリ間の利用の違い

同一カテゴリに配置されたウェブサイトの時間軸を意識した兄弟カテゴリ(横関係)の利用パターンを比較する．例えば，タレントカテゴリにおいて SMAP とモーニング娘のサイト利用され方の比較を行う．利用者に違いが有ると考えられる．

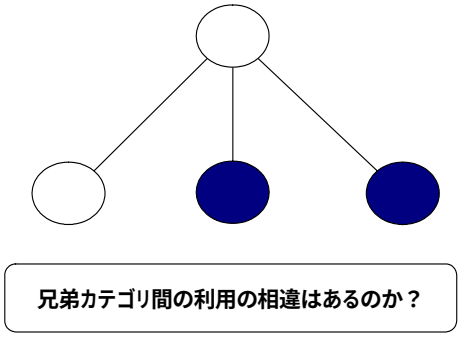


図 3. 兄弟カテゴリの利用比較

4.3. 親子カテゴリ間の利用の違い

親子(縦関係)間のカテゴリの利用の違いを調べる．兄弟間ならびに親自身の利用のされ方の重ねあわせがどう影響するかを調べる．

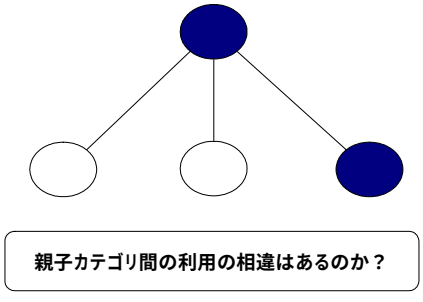


図 4. 親子カテゴリ間比較

階層型カテゴリの利用特徴を調べる上で，親子間における子の親への影響を調べることは重要である．

ある子供 i の影響度が大きい場合，x の特徴は i に支配されているといえる．この場合 x と i の利用頻度パターンは非常に類似していることになる．また大きな影響を及ぼす子供が存在しない場合，x の特徴はすべての子供を重ね合わせたものであるが，x とそれぞれの子供が類似している場合や，すべての子供を重ねあわせてはじめて x の特徴が現れるという場合が考えられる．後者では個々のウェブサイトでは分からなかった特徴がカテゴリとして扱うことで得られるものであり，たとえば，個々では周期があるようには見えなかったものが重ね合わせたカテゴリの利用頻度パターンを調べることによって周期を発見することができるといった具合である．

4.4. カテゴリ間の共起関係の発見

カテゴリ類似検索を行うことで，一見関係のないと思われるカテゴリ間で，時系列における共起関係を調べることが可能になる．例えば，スポーツのアテナオリンピックに関連するカテゴリが人気があったときに，アテネへの旅行のカテゴリが付随して人気が出るという関係が分かるといった具合である．このような共起関係が分ればより利用者に便利なサイト構成などに応用することが可能である．

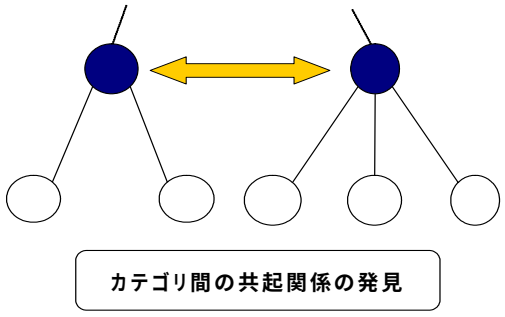


図 5．カテゴリ間の共起関係

5. 利用頻度パターンによる類似判断手法

精度の高いウェブサイト間の相関性検索(利用パターン分類)を行うために類似性を判断する必要がある．

5.1. 時間帯別利用型によるカテゴリ分類手法

カテゴリの時間帯による利用型を定義する．以下のように一日を 6 時間刻みの時間帯に分ける．

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

利用頻度において各時間帯の利用頻度を集計し全体の利用頻度に占める割合を求める．最大の割合を占める時間帯の割合が閾値を超えるとき，利用型をその時間帯に決める．これによりカテゴリを朝方利用型，昼間利用型，夜間利用型，深夜利用型に分けることができる．

5.2. フーリエ係数を用いた類似判定手法

2 つの利用頻度パターンをユークリッド距離関数と閾値をもって類似を判定することができる．波形同士を比較する際には 2 つの波の離散フーリエ変換(DFT)によって得られるフーリエ級数 {Q, T} を用いたユークリッド距離 $D(Q, T)$を用いるのが一般的である．閾値 ε が与えられ，クエリーQ に対して $D(Q, T) \leqq \varepsilon$ を満たす T すべてを類似していると判定する．

フーリエ係数を用いた類似判定の研究における速度向上の取り組みは Agrawal らの研究[7]が基礎となっている．[7]では周波数の小さい k 個のフーリエ係数を利用した下界(Lower Bound)の計算について述べている．GEMINI と呼ばれる下界距離計算とクエリーと比較する候補者検索のための多次元インデックスを利用している．

- **離散フーリエ変換**

離散系に応用したフーリエ変換が離散フーリエ変換(DFT)である．変換式は以下の通りである．

$$X(f_{K/N}) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} x(n) e^{-j\pi k \frac{2n}{M}}, k = 0,1 \ldots N-1$$

ただし，X(f)は周波数領域のデータであり x(n)は時間領域の基本周期となる N 個のデータである．$f_{k/N}$ の範囲が N/2 であるのはサンプリング定数によるものである． $X\left(f_{k/N}\right)$は複素数になるが，これは周波数成分の強さと位相を結果とするためである．周波数ごとの複素正弦関数の線形結合によって元の信号の復元が可能である．

ベクトルの強さ P をフーリエ係数の絶対値の二乗値で表現できる．

$$P(f_{k/N}) = \| X(f_{k/N}) \|^2, k = 0,1 \ldots \left\lceil \frac{N-1}{2} \right\rceil$$

P の大きなフーリエ係数は，大きなエネルギーをもったフーリエ係数であり元の波の形に大きな影響を与えている．このエネルギーの大きなフーリエ係数を用い

て周期性を強調した利用頻度パターンを復元することができる.

5.3. バーストを用いた類似カテゴリ判定手法

「短期的に起きる急激な利用頻度増加」をバーストとした．バーストは利用者の短期的に盛り上がる興味を知る上で重要である．例えば，アテネオリンピックのようなイベントがテレビで取り上げられスポーツに関連するカテゴリの利用が大きく増えるが，イベントが過ぎると利用も減るといった具合である.

バーストはカテゴリの特徴を示すものであり,カテゴリ間のバーストの類似を調べることによってカテゴリの類似性を判断することができる.

バースト検出では,利用頻度が利用平均値の一定数倍を超え，かつ直前の時刻の利用頻度が利用頻度平均値より低いときバーストとする.

各カテゴリからバーストを抽出し,次の情報を取り出す.

- バースト発生時刻
- バーストの発生回数
- 各時刻の利用頻度比率（平均を１とする）

これらの情報を用いてバーストを用いた類似判断手法を２つ提案する.

5.3.1. 共起バーストの発生回数の類似度による類似判定

カテゴリＡとカテゴリＢの共起するバーストの数の,ＡとＢのバースト総数に対する割合の平均を取る.これによりバーストという特徴点に注目した，カテゴリ間の全体の類似度を求めることができる．二つの平均を取ることによってＡとＢの共通の類似度を求めることができる.

$$Sim(A,B)=\frac{1}{2}\left(\frac{NB_{A\cap B}}{NB_A}+\frac{NB_{A\cap B}}{NB_B}\right)$$

NB:バーストの発生回数

$NB_{A\cap B}$:ＡとＢの共起バースト 回数

5.3.2. 共起バーストの利用頻度加算による類似スコアによる類似判定

共起バーストの利用頻度の規模を考慮した類似判定を行う．カテゴリＡとＢの共起バーストの各利用頻度のうち共通部分を加算することによってバーストの類似スコアを計算する．共起回数だけではなく二つのカテゴリの共有利用頻度比率を用いることによってカテゴリの利用頻度パターンの類似を求める.

$$Sim\_score(A,B)=\sum_{t=1}^{n}\min(b_t(A),b_t(B))$$

t:ＡとＢの共起バースト発生する時刻

n:共起バーストの数

$b_t(A)$:Aの時刻tのバーストの利用頻度

$b_t(B)$:Bの時刻tのバーストの利用頻度

6. 実験データおよび実験

本研究では Yahoo!カテゴリというディレクトリ型検索エンジンのカテゴリの階層構造に注目し，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，カテゴリの親子間の利用頻度への影響度，兄弟カテゴリの利用の違いについても調べた．解析には Perl を利用し，カテゴリ構造抽出とテキストタイプのプロキシログ処理行った.

6.1. 実験データ

Yahoo!カテゴリは“エンターテイメント”や“メディアとニュース”といった主要なカテゴリ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カテゴリ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カテゴリ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．カテゴリ情報を使うことによってカテゴリの抽出，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．カテゴリページに含まれる主要な情報として次の 2 つがある.

- Yahoo!カテゴリ
  子のカテゴリへのリンク.「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.
- Yahoo!登録サイト
  そのカテゴリに関するサイトで”yahoo”をアドレスに含まないもの.

Yahoo!カテゴリにおいてリンク先は他のカテゴリ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．カテゴリ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.

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．抽出したカテゴリ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 04/07/01 から 04/10/31 の 4 ヶ月分である． Kyoto I-net では 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．実験に利用するレコード総数は 148,467,531 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 19,891,523 であり全体の 13.4%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．

なお Kyoto I-net からは個人が特定できる情報は一切削除された形式でアクセスログの提供を受けたため，Kyoto I-net の会員のプライバシーを侵害することはない．

アクセス履歴に現れた Yahoo!カテゴリと登録サイトは 70,655 種であった．利用頻度パターンの類似分類を行う場合には 2004 年 9 月 1 日～10 月 31 日間のアクセス履歴を利用し，もっとも精度の高い 1 時間単位の集計(hour)データを用いた．グラフの中にはゼロがほとんどで特徴を見出せないものも多かったため，集計した各カテゴリの持つ最大利用頻度(MAX)によってフィルタリングを行った(表 1 参照)．

MAX	MAX≧10	MAX≧100
カテゴリ数	12,684	705

表 1．最大利用頻度(MAX)によるフィルタリング

予備実験では MAX≧100 を満たす 705 種のカテゴリを対象として行うことにした．

6.2. 実験および評価

提案する類似判定手法を用いた利用者の興味分析を行った．

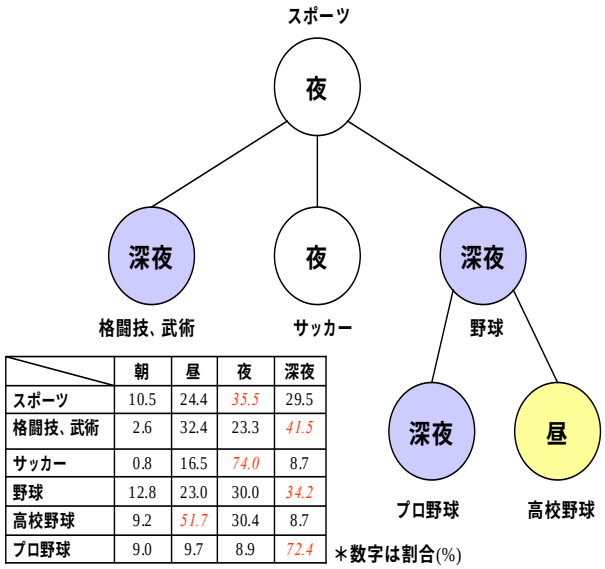
6.2.1. 時間帯別利用型の調査

705 個のカテゴリの時間帯別の利用型の分類を行う．利用型を判断する際に用いる閾値を 0.3 と 0.4 にしたときの結果を 6.4 に示す．0.3 においてはウェブの一般的な利用スタイルから夜間利用型のものが最大である．4 つの利用型によく分類できているといえる．昼間利用型の中にはディズニー映画のサイト，本屋のサイトなどがあった．

閾値/時間帯	朝	昼	夜	深夜	未分類	合計
0.3	17	158	292	229	9	705
0.4	16	96	140	143	310	705

表 2．時間帯別の利用型への分類

階層カテゴリにおける時間帯別の利用型の違いを調べるために閾値 0.3 で行った利用型分類結果の中から“趣味とスポーツ＞スポーツ”カテゴリを親とした階層構造を図 6 に示す．図に表れるカテゴリはフィルタリングの結果残ったものである．“趣味とスポーツ”カテゴリの下には 20 個以上の子カテゴリが配置されている．図を見ると親子関係，兄弟関係にあるカテゴリでも利用型に違いがあることが分かる．“サッカー”の利用は 74.0%が夜間で典型的な夜間利用型である．“プロ野球”は 72.4%が深夜利用である．一方で“高校野球”は 51.6%が昼間の利用であり，利用の違いがあることが分かる．格闘技，武術”は 72.4%で深夜利用型であった．時間帯型利用型は親子で同じになるとは限らず，“スポーツ”という同一カテゴリにおいても利用型が違うものがあることが分かる．



	朝	昼	夜	深夜
スポーツ	10.5	24.4	*35.5*	29.5
格闘技、武術	2.6	32.4	23.3	*41.5*
サッカー	0.8	16.5	*74.0*	8.7
野球	12.8	23.0	30.0	*34.2*
高校野球	9.2	*51.7*	30.4	8.7
プロ野球	9.0	9.7	8.9	*72.4*

＊数字は割合(%)

図 6.“スポーツ”カテゴリの時間帯別利用型

6.2.2. 実験 2:共起カテゴリ発見

ここでは階層構造上関係性が低いカテゴリ間の共起関係を調べる．距離関数を用いた類似判断とバーストを用いた類似判定手法によって得られる相関性の強い上位 10 個の関係から提案した各手法の特長について報告する．ランキングにおいてはスペースの関係からカテゴリの階層構造を途中省略して記述した部分があるが，評価のうえで問題はない．(2 回目以降の出現李は斜体表示している．)

A.フーリエ係数を用いた類似判定

表 3 を見てみると意味の近い関係は見受けられない．これは利用の時刻が類似したことによって得られたものと考えられる．図 7 は一位の関係の利用頻度パターンである．利用頻度が少なく，利用頻度がゼロの時刻が多いことが分かる．このような場合，類似距離の計算ではフーリエ係数を求める必要は無く，利用頻度の値を用いてユークリッド距離を求めれば十分である．

	カテゴリ	カテゴリ
1	メディアとニュース＞ビジネスと経済	地域情報＞＞＞鹿児島
2	コンピュータとインターネット＞情報と資料＞辞書	健康と医学＞＞産婦人科＞病院
3	ビジネスと経済＞＞本＞＞絵本＞クレヨンハウス	地域情報＞＞＞大阪＞枚方
4	*辞書*	地域情報＞＞＞愛知＞＞ビジネスと経済
5	*メディアとニュース＞ビジネスと経済*＞新聞	*鹿児島*
6	*辞書*	*産婦人科*
7	ビジネスと経済＞＞＞中央銀行	*ビジネスと経済＞＞本＞＞絵本*
8	*中央銀行*	*産婦人科*
9	コンピュータとインターネット＞＞OS＞WindowsCE	*大阪＞枚方*
10	*WndowsCE*	地域情報＞＞＞＞鈴鹿＞＞カーレース

表 3. フーリエ係数を用いたら類似距離による共起カテゴリ上位 10

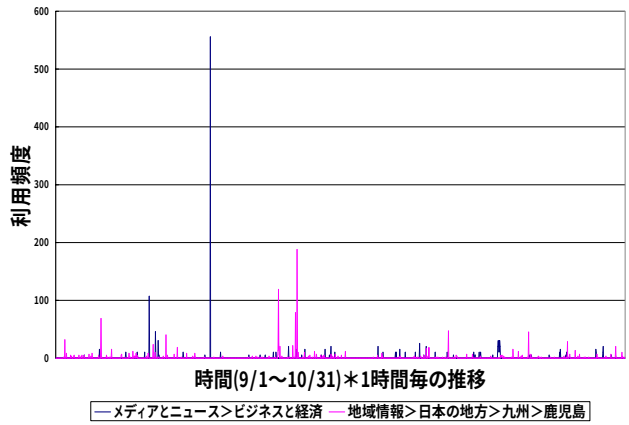


図 7. フーリエ係数を用いた距離関数より得られた 1 位の共起カテゴリの利用頻度パターン

B.共起バーストの発生回数の類似度を用いた類似判定手法

1 位には“コンピュータとインターネット＞無料サービス＞無料メール”と“ビジネスと経済＞ショッピングとサービス＞インターネット・サービス＞エキサイト”という無料メールとポータルサイトのエキサイトが共起していることが分かる．どちらもインターネットに関するカテゴリであり利用に関係があると思われる．6 位の共起カテゴリは“エンターテイメント＞芸能人，タレント”と“ビジネスと経済＞企業間取引＞エンターテイメント＞タレント，スタッフ”であった．“エンターテイメント＞タレント“という同じトピックを扱うカテゴリが共起していることからあきらかに意味的に近い共起関係が抽出できたことになる．8 位の共起カテゴリでは”芸能と人文＞デザインアート“と”ビジネスと経済＞ファッションとサービス＞服飾，ファッション“がある．これも意味的に関係のある共起関係であるといえる．

	カテゴリ	カテゴリ
1	コンピュータとインターネット＞＞無料メール	ビジネスと経済＞＞＞ポータルサイト＞エキサイト
2	ビジネスと経済＞ショッピングとサービス＞＞手芸＞ビーズ	趣味とスポーツ＞＞＞プロ野球＞メディアとニュース
3	*ビーズ*	趣味とスポーツ＞＞＞プロ野球＞メディアとニュース
4	*ビーズ*	趣味とスポーツ＞＞＞プロ野球＞メディアとニュース
5	*ビーズ*	趣味とスポーツ＞＞＞プロ野球＞メディアとニュース
6	エンターテイメント＞芸能人	ビジネスと経済＞＞エンターテイメント＞タレント
7	エンターテイメント＞芸能人	ビジネスと経済＞＞＞タレント＞芸能プロダクション
8	芸術と人文＞デザインアート	ビジネスと経済＞ショッピングとサービス＞ファッション
9	地域情報＞＞＞愛知＞市町村	ビジネスと経済＞＞＞ソフトウェア＞ゲーム
10	芸術と人文＞デザインアート＞ファッション	ビジネスと経済＞買い物＞ファッション

表 4. 類似度による共起カテゴリ上位 10

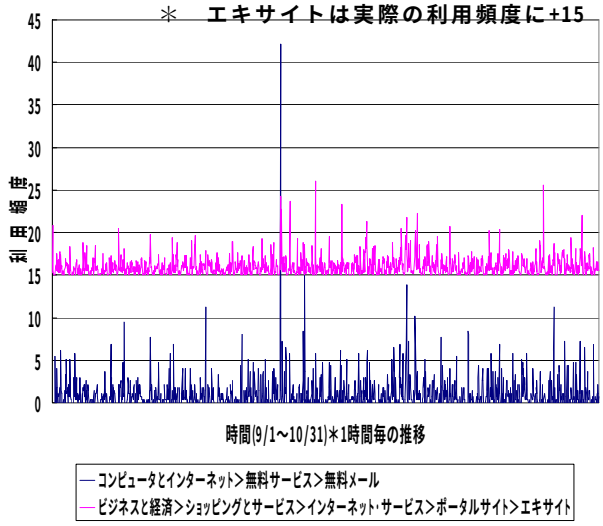


図 8.類似度 1 位の共起カテゴリ利用頻度パターン

これらの共起関係は利用頻度を無視しているこ

とから，利用頻度の低いカテゴリの共起関係をみつけることもできるというメリットがある．一方でより利用頻度の大きい，つまり人気のある共起関係を優先して検出することはできない．

C.共起バーストの利用頻度加算による類似スコア計算を用いた類似判定

	カテゴリ	カテゴリ
1	メディアとニュース＞テレビ＞番組	ビジネスと経済＞＞メディアとニュース＞テレビ局＞テレビ朝日
2	番組	*テレビ局*＞日本テレビ
3	*番組＞アクション＞仮面ライダー*	*テレビ朝日*
4	*場組み＞アクション*	*テレビ朝日*
5	*健康と医学*	ビジネスと経済＞＞出版＞小学館
6	健康と医学＞病院	*出版*
7	ビジネスと経済＞買物とサービス＞＞バイク	地域情報＞＞＞兵庫＞＞ 神戸
8	健康と医学	*神戸*
9	地域情報＞＞＞愛知＞市町村	*出版*
10	芸術と人文＞デザインアート＞ファッション	各種情報と情報源

表 5. 類似スコアによる共起カテゴリ上位 10

＊ テレビ朝日は実際の利用頻度に+15

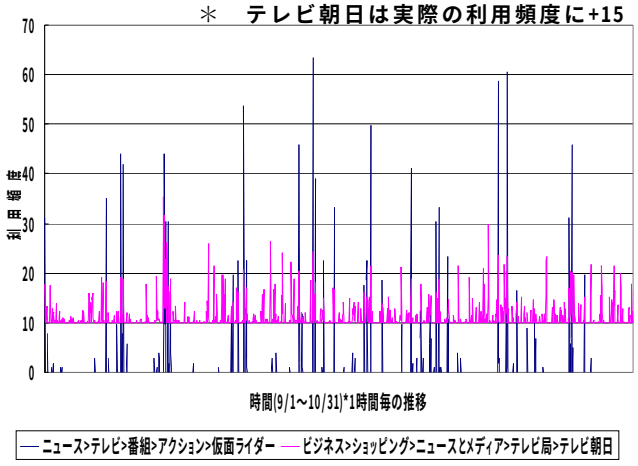


表 9.意味の近い共起関係
（3 位の共起カテゴリの利用頻度パターン）

バーストの利用頻度の大きさを考慮することにより意味の近く，かつ利用の大きい関係が見つけられると考えられる．1 位と 2 位の共起カテゴリはテレビ番組とテレビ局の関係が強いことを示している．3 位の仮面ライダーの番組はテレビ朝日で放映されており，その仮面ライダーを調べるためにテレビ朝日を訪れる関係があると考えられる．両カテゴリのウェブページではお互いのウェブページへの直接のリンクが準備されていないことから両カテゴリを直接リンクでむすぶことで利用者にとって使いやすいカテゴリ構造を構成することができる．このような関係を多く見つけることによってカテゴリの再構成を行うことができると考えられる．

7. まとめ

本稿では利用者のウェブ上の行動を記録したアクセス履歴と WWW のインデックスであるディレクトリ型検索エンジンを用いてウェブサイトのアクセス履歴の相関性を調べる方法について述べた．

時間帯別利用型によるカテゴリ類似判定によるカテゴリの特徴分析を行った．また共起カテゴリ発見においてはフーリエ係数とバーストを利用した手法を提案し検証した．今後は類似判定に用いる閾値について検討する必要がある．本稿で提案した手法はネットワークトラフィックなどのアクセス履歴解析に利用することができる．

文　献

[1] C. Wang and X. S. Wang “Multilevel filtering for high dimensional nearest neighbor search,” In ACM SIGMOD Workshop on Research issues in Data Mining and Knowledge Discovery, 2000

[2] D. Rafiei and A. Mendelzon “Efficient retrieval of similar time sequence using dft,” In Proceedings of FODO, 1998

[3] 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

[4] M. Vlachos, C. Meek and Z. Vagena “ Identifying Similarities, Periodicities, and Bursts for Online Search Queries, ” In Proceedings of SIGMOD, 2004

[5] M. L. Hetland, “A SURVEY OF RECENT METHODS FOR EFFICIENT RETRIEVAL OF SIMILAR TIME SEQUENCES,” Data Mining In Time Series Databases (Series in Machine Perception and Artificial Intelligence), ISBN: 9812382909, pp23-40

[6] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger “The r*-tree: An efficient and robust access method for points and rectangles,” In proceedings of ACM SIGMOD, 1990

[7] R. Agrawal, C. Faloutsos, and A. Swami “Efficient Similarity Search in Sequence Databases,” In Proceedings of the 4th FODO, pp69-84, 1993

[8] DMOZ(http://dmoz.org/)

[9] Google(http://www.google.co.jp)

[10] MSN(http://www.msn.com)

[11] Yahoo!カテゴリ(http://dir.yahoo.co.jp)

[12] 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