DEIM Forum 2018 E7-4

不連続なセッションで構成されたウェブアクセスログデータを用いたコンテンツ間の関連性の分析

鈴木　亮平† 伊藤　雅博†† 廣田　雅春††† 荒木　徹也†††† 石川　博††††

† 首都大学東京　システムデザイン学部　〒 191–0065 東京都日野市旭が丘 6-6
†† NCS-P 研究所
††† 岡山理科大学総合情報学部　〒 700–0005 岡山県岡山市北区理大町 1-1
†††† 首都大学東京大学院　システムデザイン研究科　〒 191–0065 東京都日野市旭が丘 6-6
E-mail: †suzuki-ryouhei2@ed.tmu.ac.jp, ††masahiro-ito@ncs-pro.com, †††hirota@mis.ous.ac.jp,
††††{araki,ishikawa-hiroshi}@tmu.ac.jp

あらまし　近年，CTR などを向上させるためのユーザの閲覧履歴を用いた Web サイトの構造改善の研究が盛んに行われている．Web サイトの構造を向上させる研究では，ユーザごとの長期間に渡る閲覧履歴が用いられることが多い．しかし，ユーザをアカウントなどで制御せずに，Cookie などを含むセッションが一時的にユーザを識別するのみのウェブサイトも存在する．そこで，本研究では，アクセスログデータに含まれる 1 つのセッションを 1 人のユーザとみなすことで，擬似的にユーザを区別して閲覧履歴を分析することで，ウェブサイトの構造改善に取り組む．本研究では，それぞれのセッションにおけるページ移動を元に特徴量を作成し，クラスタリングを適用する．クラスタリングの結果から，サイト構造を改善するための，関連性の高いウェブページの組み合わせを抽出し，ウェブページ間の関連性を分析する．

キーワード　アクセスログ分析，Web サイト，ウェブログデータ

1. はじめに

パーソナルコンピュータ及びスマートフォンの普及により，ウェブサイトを日常的に閲覧する手段が増え，ますますウェブサイトの重要性が高まっている．それに従ってコンテンツを提供するウェブサイトの需要も高まり，非常に多くの Web サイトが日々作成，更新されている．ウェブサイトの中にはマーケティングや広報として利用されているものも多くあり，それらをはじめとする多くのウェブサイトでは，より多くのユーザに利用されることが理想とされる．より多くのユーザに利用してもらうためにはユーザのニーズに即したウェブサイトを作成する必要がある．ここでのニーズとは，コンテンツの内容だけでなく，サイトのユーザビリティも含まれている．

一般的にサイト作成者は，サイト閲覧ユーザのアクセスログデータを得ることができる．アクセスログデータには Cookie, IP アドレス，端末情報，ブラウザ情報などが含まれている．このようなアクセスログデータから，現状のウェブサイトがユーザのニーズに合っているのかを見極めることは，サイト運営において必須となる．一般的にアクセスログデータ解析ではページ別アクセス頻度や CTR(Click Through Rate)，時間別のアクセス頻度，当該ウェブサイトへのリンク元の集計，OS やブラウザ別の集計などが行われる．そのため，各々のユーザのコンテンツへの興味・関心を考慮したサイト構造の改善には向いていない．

各々のユーザのニーズを満たすための手法として，ウェブページ推薦システムがある．はじめに，登録されたユーザ ID または Cookie や IP アドレス等を用いて同一ユーザのアクセスを継続的に追跡する．次に，追跡によって得られた行動パターンを解析し，分類することによって，グループ毎に，ニーズに合わせたページ (コンテンツ) を次のアクセス先として推薦することを可能としている．しかし，これは長期的に同一であると判断できるユーザを追跡することによって成り立っている．ユーザ ID が存在していない場合，Cookie や IP アドレス等を用いて以前にアクセスしたことのある利用者か判断することになる．現在一般的に用いられているスマートフォンでは，移動しながらの利用も想定される．その場合，移動に合わせて接続する基地局が変更され，基地局が変わると IP アドレスも更新される．そのため，同一ユーザの抽出には適さない．Cookie に関しても，自発的に Cookie を定期的に削除するユーザや，シークレットブラウザ (ページを閉じるごとに Cookie を自動削除する機能を持つブラウザ) を使用するユーザが増えており，Cookie を用いて長期的に同一ユーザを捉えることは難しい．

本研究では，ユーザ ID が存在しないアクセスログデータにおいて，短期的なユーザの行動から，ユーザに沿ったサイト構造改善のために，ユーザのニーズを抽出する手法を提案する．具体的には，アクセスから切断までの一連の行動を 1 セッションと定義し，アクセスログデータに含まれる 1 つのセッションを 1 人のユーザとみなして閲覧履歴をもとに分析を行う．シークレットブラウザや手動による Cookie の削除は通常閲覧前か閲覧後に行われるため，同一セッションにおいては Cookie が変化しないと推察される．よって，ユーザ ID が存在しないアクセスログデータにおいてもセッションの抽出は可能であると

考えられる．まず始めに，セッションにおけるページ移動を特徴量としてベクトル化し，セッションのクラスタリングを行う．クラスタリングによって生成された各セッション群に対して出現頻度が高いコンテンツを抽出し，コンテンツ間の関連性を分析する．

本論文の構成は以下の通りとなる．2 章では，アクセスログデータに関する関連研究を述べる．3 章では，セッションを用いたクラスタリング及びコンテンツの抽出手法について述べる．4 章では，実験の結果と考察を述べる．5 章では，本研究のまとめと今後の課題について述べる．

2. 関 連 研 究

Web アクセスログデータを用いた研究は数多く存在する．

2.1 アクセスログデータへのアプローチ

木虎ら [1] は，個人の Web 上での行動も嗜好・価値観に基づいていると考え，Web アクセス履歴を元に個人の閲覧したページを解析し，その価値観を類推することを試みている．

三原ら [2] は，Web アクセスログ解析においてはページ遷移だけでなくページ閲覧時間も重要な要素と考えた．同じページでもユーザごと，あるいは経路ごとにそのページのユーザの興味の度合いは異なり，その違いは閲覧時間の長さから推定できると考えた．ページの閲覧時間に重み付けをほどこしたセッション毎のグラフデータベースを作成し，グラフマイニングによる Web アクセスパターンを抽出する Web マイニング手法を提案している．そして実際に閲覧時間に基づくユーザの興味を踏まえたパターン抽出が可能であることを実験で示している．

大塚ら [3] は，統計的に偏りなく抽出された人 (パネル) を対象として収集された URL 履歴 (パネルログ) を対象とした研究を行っている．類似した Web ページを抽出するパネルログ解析システムを提案し，URL を基にした解析では捉えがたい大規模なユーザの行動パターンを抽出している．

2.2 アクセスログデータのセッションに関する研究

山元ら [4] は，Web アクセスログ解析によって抽出した LCS(Longest Common Subsequences) を用い，ユーザの過去のアクセス行動からそれに続くアクセスページを推薦する手法である WRAPL について解析を行った．実際の Web アクセスログを用いた実験を通して， WRAPL の効果を詳細に解析した．その後，推薦ページの優先順位付け方法を改良した WRAPL-FLP 法を提案し，それによって推薦精度が向上することを確認している．

鈴木ら [5] は，ユーザの経路が一意に決まるのではなく確率的に推移するという遷移確率の考え方のもと，遷移確率情報を付加したセッション同定について研究を行い，それらを可視化するシステムを提案している．

2.3 ユーザへのページ推薦に関する研究

高須賀ら [6] は，収集した閲覧履歴をもとに，閲覧行動が類似するユーザの履歴からユーザへ Web ページを推薦するシステムを実装した．

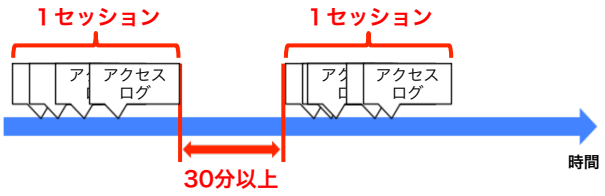

図 1 セッション

山口ら [7] は，Web 広告推薦をより効果的にすることを目的として，従来の FQDN(Fully Qualified Domain Name) によらない，閲覧した Web ページのカテゴリ情報を用いたユーザの潜在的興味分析方式について提案を行った．FQDN とは，ホスト名，ドメイン名（サブドメイン名）を省略せずに指定した記述形式のことである．FQDN は異なるがページ内容 (カテゴリ) が類似した Web サイトや，同一の FQDN でもページの内容 (カテゴリ) が異なる Web サイトを閲覧しているユーザ間の類似度を適切に判定することができないという問題点が挙げられる．

3. 提 案 手 法

本章では，セッションを用いたクラスタリング及びコンテンツの抽出手法について述べる．

3.1 セッションの抽出

本節では，アクセスログデータからセッションを抽出する方法について述べる．一般的なアクセスログデータには，IP アドレス，アクセス日時，現在閲覧している URL(遷移前 URL と定義する)，次にアクセスしようとリクエストしている URL(遷移後 URL と定義する)，Cookie などが含まれている．本稿では，ログのアクセス日時と Cookie を用いて，ユーザのセッションを抽出する．同一 Cookie をもつアクセスログデータは同一ユーザのものであるため，はじめに，Cookie ごとのログデータを収集する．このとき，有効な Cookie をもたないログデータが存在するため，このようなログデータは本論文では用いていない．本研究では，ユーザのサイトへのアクセスから切断までのログをセッションとして扱う．Cookie ごとのログデータより，あるログから 30 分以内に次のアクセスが存在するログ群を，ユーザのセッションとして抽出する．また，同一 Cookie においても，次のアクセスまで 30 分以上が経過したものは，別のセッションとして扱う．図 3.1 にセッションの例を示す．

3.2 セッションに対する特徴量の作成

本節では，抽出したセッションに対して特徴量を付与する方法について述べる．本論文で用いたニュースサイトのアクセスログデータには，URL ごとにニュース記事のジャンルが存在している．そこで，抽出したセッションのアクセスログに対して，URL をもとにニュース記事のジャンルを付与する．このとき，ひとつのログには，遷移前 URL と遷移後 URL が存在するた

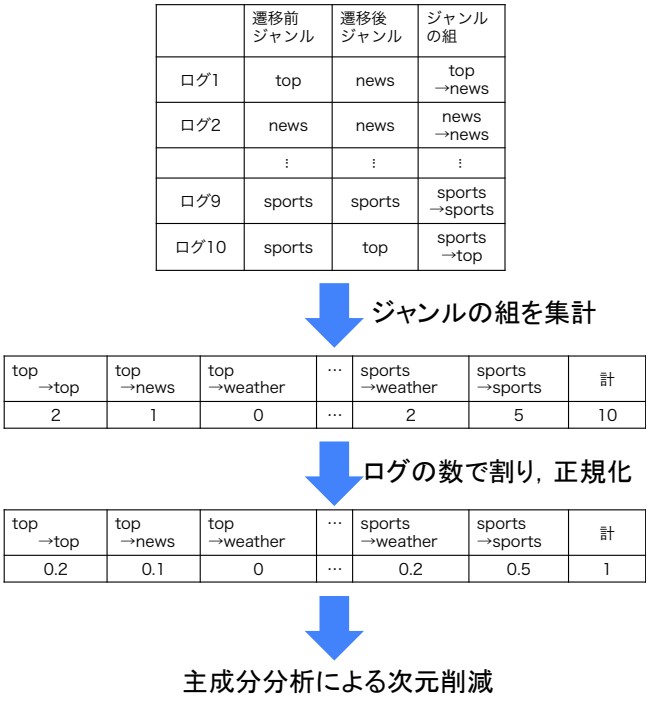

図 2　あるセッションにおける特徴ベクトルの作成手順

め，これをユーザの閲覧記事の遷移とみなすことができる．そこで，遷移前 URL のニュースジャンルと遷移後 URL のニュースジャンルの組をベクトルの一つの要素として，セッション内での遷移回数をもとに，特徴量を作成する．セッションはそれぞれ含まれるログの数が異なるため，全要素をログの数で割り，正規化を行う．図 2 にセッションの特徴ベクトルの作成手順を示す．また，これらのベクトルの次元数は，ニュースジャンルの数の二乗となる．このままでは要素数に対して次元数が非常に大きくなってしまうため，次元削減を行う．次元削減には，主成分分析 (PCA: Principal Component Analysis) を用いる．

3.3　セッションのクラスタリング

本節では，セッションごとに作成した特徴ベクトルを用いたセッションのクラスタリングについて述べる．今回，クラスタリングは k-means++法を用いた．k-means 法は代表的な非階層型クラスタリング手法のひとつであり，それぞれのクラスタに属する各ノードとクラスタ重心との距離の総和を最小化することでクラスタ分割を行う．k-means++は各ノードとクラスタ重心との距離の総和に重みをつけて計算することで，k-means におけるランダムに選ばれる各ノードの初期値によってクラスタリング結果が変化してしまう点を解消したものである．クラスタリングに用いる距離は任意であるが，ここでは k-means における一般的な距離計算方法であるユークリッド距離を用いる．k-means でのクラスタ数決定についてはクラスタ内誤差平方和を参考に決定する．

3.4　クラスタ毎に特徴を抽出，分析

本節では，クラスタリング結果をもとに，各クラスタの分析方法について述べる．

表 1　アクセスログデータの詳細例

IP アドレス	アクセス時間	リクエスト URL
49.96.19.177	2016-04-01 09:30:03	GET or POST /(URL)
遷移前 URL	使用ブラウザ	Cookie
(URL)	Mozilla/5.0(Linux; U;)	nscookie=(任意の文字列)

クラスタリングを行うことにより，同一クラスタ内では同一の遷移を行なっているセッションがまとまっていると考えられる．同一の遷移を行なっているセッションは類似した利用目的を持ったユーザのセッションのまとまりと推察される．各クラスタごとにジャンルの出現数を集計する．集計した遷移を用いて，各遷移に対して，クラスタ全体の遷移を母数とした割合を算出する．その割合が他のクラスタと比較して高ければ，クラスタにおける特徴と考えることができる．こうして得られた複数の特徴はユーザが似通った利用目的で利用していると考えられる．その仮定に基づき，実際にその組み合わせの移動があるかどうか検証・考察する．

4.　実　　験

本章では，提案手法を適用した結果について考察する．

4.1　データセット

本実験のデータセットは，スマートフォン向けサービスとして運営されているニュースサイトのアクセスログデータを用いる．表 1 はアクセスログの例を示したものである．期間は 2016 年の 4 月 1 日から 30 日までの 1ヶ月，アクセスログデータ数は 3,691,983 件で，そのうち本論文で対象とする Cookie を保持しているデータ数が 2,151,104 件であった．提案手法によって 320,519 セッションを取得した．

4.2　特徴量の作成，クラスタリング

本節では，セッションごとの特徴量作成及び，クラスタリングの結果について述べる．URL 毎にニュース記事のジャンル付与した結果，遷移前 URL をもとに付与されたジャンルは 88 種類，遷移後 URL をもとに付与されたジャンルは 85 種類であった．これらのジャンルをもとに，3.2 節の手法を用いて特徴量を作成する．今回，主成分分析には Python の機械学習パッケージである scikit-learn [8] を用いた．主成分分析の結果，このときの次元数は，累積寄与率が 0.8 を超えた 8 次元を用いた．クラスタリングにおいても同じく scikit-learn [8] を用いた．クラスタ内誤差平方和の減少率が 10 %以下になる時点である 7 をクラスタ数とした．表 2 に各クラスタのセッション数及びセッションに含まれるログの合計数を示す．

4.3　クラスタ毎に特徴の集計，コンテンツの組み合わせの抽出

各クラスタでの遷移後のジャンルの集計結果の上位 5 件を表 3 に示す．表 3 より，クラスタ内での出現数をみると，複数のクラスタにおいて，コラム (特集) が上位に出現している．そこで，コラム (特集) の出現数に対しての割合で考える．すると，全コラム (特集) のうち約 42 %がクラスタ 5 に含まれている．これは全クラスタにおいて一番高い割合である．また，ク

表 2　クラスタリング結果

	セッション数	ログの合計数
クラスタ 1	74,104	361,590
クラスタ 2	59,522	361,392
クラスタ 3	39,796	248,061
クラスタ 4	18,520	40,223
クラスタ 5	44,639	466,734
クラスタ 6	19,249	122,114
クラスタ 7	64,689	550,990

表 3　各クラスタでの遷移後ジャンル上位 5 件の出現件数及び割合

	1 位	2 位	3 位	4 位	5 位
	ニュース	総合トップ	コラム (特集)	社会 (主要ニュース)	もっと見る
クラスタ 1	136,034(21 %)	108,026(23 %)	24,237(22 %)	16,997(26 %)	15,812(27 %)
	スポーツ	スポーツトップ	総合トップ	ニュース	天気トップ
クラスタ 2	125,972(35 %)	45,856(13 %)	43,149(12 %)	35,408(10 %)	27,380(8 %)
	総合トップ	ニュース	もっと見る	コラム (特集)	スポーツトップ
クラスタ 3	90,516(36 %)	84,901(34 %)	14,799(6 %)	8,882(4 %)	7,491(3 %)
	総合トップ	ニュース	スポーツ	コラム (特集)	社会 (主要ニュース)
クラスタ 4	25,651(64 %)	3,851(10 %)	991(2 %)	734(2 %)	725(2 %)
	ニュース	総合トップ	コラム (特集)	社会 (主要ニュース)	もっとみる
クラスタ 5	208,816(44 %)	74,881(16 %)	45,483(10 %)	25,362(5 %)	19,023 (4 %)
	天気	ニュース	総合トップ	天気トップ	天気設定
クラスタ 6	78,454(64 %)	7,425(6 %)	7,016(6 %)	6,994(6 %)	2,551(2 %)
	ニュース	総合トップ	スポーツ	天気トップ	天気
クラスタ 7	174,657(27 %)	113,601(20 %)	46,780(8 %)	44,255(8 %)	37,048(7 %)

ラスタ 1 は次に割合の高い約 22 %である．これより「コラム (特集)」をクラスタ 5 における特徴とする．同様にして,「社会 (主要ニュース)」もクラスタ 5 に最も多く出現している．クラスタ 2 では「スポーツ」が 69 %,「スポーツトップ」が 53 %含まれており，全クラスタ中で最も多く出現しているので，特徴と考えられる．クラスタ 3，4 では，最も含まれているジャンルが存在しなかった．クラスタ 6 では「天気」が最も多く出現している．クラスタ 7 では「総合トップ」が 24 %,「天気トップ」が 49 %含まれており，全クラスタ中で最も多く出現しているので，特徴と考えられる．クラスタ 2，5，7 から得られた組み合わせは，ユーザのニーズにおいて関連性の高いコンテンツの組である可能性が考えられる．

4.4　連続した移動の判定

次に，関連性が高いコンテンツの組での連続した移動があるかを判定する．今回は基準となる片方のジャンルを持つログから前後 5 ログ以内にもう一方のジャンルを持つログが出現した場合をジャンル間の連続した移動とする．クラスタ 5 において遷移後のジャンルとして「社会 (主要ニュース)」を含む，すなわち閲覧したセッション数は 14,272 件,「コラム (特集)」を含むセッション数は 14,604 件,「社会 (主要ニュース)」と「コラム (特集)」をどちらも含むセッションは 5,614 件であった．そのうち，二つのジャンル間を連続して移動しているセッションは 4,854 件であった．これはどちらも含むセッションのうち 86 %であった．

4.5　考　　察

本節では，クラスタリング結果及びコンテンツの組み合わせの抽出，連続した移動の判定それぞれの結果に対する考察を述べる．本論文ではクラスタリング手法として非階層クラスタ分析手法である k-means++を用いた．その結果，あるクラスタにおいて特徴となる組み合わせを抽出することができ，一部では，期待する結果を得られた．しかし，他のクラスタに比べて 1/3 以下のデータ数となってしまっているクラスタも存在する．データ数の差が適当なものかの検証が必要である．また，クラスタ 1 とクラスタ 5 において遷移後ジャンル上位 5 件が重複していた．これは本来同一クラスタに属するべきだった可能性が考えられる．クラスタ数の決定方法，クラスタリング手法そのものの再検討が必要である．また，クラスタ 5 において，出現率が高い 2 ジャンルの組み合わせを関連性の高いコンテンツの組み合わせとしたが，組み合わせは 2 ジャンルのみとは限らず，3 つ以上のジャンルの組み合わせの抽出も必要である．連続した移動の判定結果としては，クラスタ 5 で抽出した関連性の高かった 2 つのコンテンツが同セッションで閲覧されている場合，86 %の割合で連続して閲覧していると判定された．これは十分に高い結果であり，クラスタ 5 において「コラム (特集)」と「社会 (主要ニュース)」を同時に閲覧するユーザが多い傾向にあると考えられる．

5.　おわりに

本論文では，サイト構造を改善するための，関連性の高いジャンルの組み合わせを抽出した．継続的にユーザを追跡できない場合でも，セッション毎に特徴を作成し，クラスタリングを行うことで，ユーザが同一セッション内で同時に見ることが多い記事のジャンルの組み合わせを抽出することができた

本論文では，Cookie を持たないアクセスログデータは除外したため，Cookie を持たない場合でも IP アドレスやブラウザ情報を用いてセッションを抽出する必要があると考えられる．また，今回はクラスタリングに kmeans++，ユークリッド距離を用いたが，それ以外のクラスタリング手法を用いた場合との妥当性の比較が課題の一つである．また，今回はクラスタ内でのジャンルの出現数を集計し，遷移数が多いジャンル同士を関連性が高いものとしたが，遷移数の多くないコンテンツからみて遷移数の多いコンテンツとの共起が高くなる可能性が考慮できておらず，今後の課題の一つである．

6.　謝　　辞

本研究は，首都大学東京傾斜的研究 (全学分) 学長裁量枠戦略的研究プロジェクト戦略的研究支援枠「ソーシャルビッグデータの分析・応用のための学術基盤の研究」及び JSPS 科研費 16K00157, 16K16158 による．

文　　献

[1] 木虎直樹, 久保征人. Web アクセス履歴に基づくユーザの価値観の類推. 人工知能学会全国大会 (第 27 回) JSAI2013, 2013.
[2] 三原宏一朗, 寺邊正大, 橋本和夫. ページ閲覧時間を考慮した web ログマイニング手法の提案. 情報処理学会研究報告知能と複雑系 (ICS), Vol. 2007, No. 67 (2007-ICS-148), pp. 39–44, 2007.
[3] 大塚真吾, 喜連川優. Web アクセスログとその利活用. 人工知能学会誌, Vol. 21, No. 4, pp. 410–415, 2006.
[4] 山元理絵, 小林大, 吉原朋宏, 小林隆志, 横田治夫. アクセスロ

グに基づく web ページ推薦における lcs の利用とその解析. 情報処理学会論文誌データベース (TOD), Vol. 48, No. SIG11 (TOD34), pp. 38–48, 2007.

[5] 鈴木康之, 木村マサオミ. Web アクセスログにおけるユーザ行動の分析. 第 69 回全国大会講演論文集, Vol. 2007, No. 1, pp. 513–514, 2007.

[6] 高須賀清隆, 丸山一貴, 寺田実. 閲覧履歴を利用した協調フィルタリングによる web ページ推薦とその評価. 情報処理学会研究報告データベースシステム (DBS), Vol. 2007, No. 65 (2007-DBS-143), pp. 115–120, 2007.

[7] 山口由莉子, 森下民平, 稲垣陽一, 中本レン, 張建偉, 青井順一, 中島伸介. Web 広告推薦のための閲覧カテゴリ情報を用いたユーザの潜在的興味分析方式. 第 9 回 データ工学と情報マネジメントに関するフォーラム, pp.B2-3, 2017.

[8] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, Vol. 12, No. Oct, pp. 2825–2830, 2011.

[9] 中嶋俊治, 中村健二, 小柳滋. 大規模 web サイトにおける web アクセスログの最長共通部分列を用いた推薦の高速化手法. 電子情報通信学会論文誌 D, Vol. 96, No. 5, pp. 1235–1245, 2013.

[10] 宇根田純治, 横田治夫. Web ログの共通シーケンス解析. 電子情報通信学会技術研究報告. DE, データ工学, Vol. 102, No. 64, pp. 7–12, 2002.

[11] 川口銀河, 田行里衣, 小林史弥. ネットワークアクセスログによる web 利用待ち時間推定 (コミュニケーションクオリティ). 電子情報通信学会技術研究報告= IEICE technical report: 信学技報, Vol. 115, No. 496, pp. 153–158, 2016.

[12] Hua-Jun Zeng, Zheng Chen, and Wei-Ying Ma. A unified framework for clustering heterogeneous web objects. In *Web Information Systems Engineering, 2002. WISE 2002. Proceedings of the Third International Conference on*, pp. 161–170. IEEE, 2002.

[13] 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ウェブ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 (TOD), Vol. 44, No. SIG18 (TOD20), pp. 32–44, 2003.

[14] 戸田誠二, 横田治夫. Web ログの lcs 解析におけるスケーラビリティ向上手法の評価. *DBSJ Letters*, Vol. 2, No. 3, pp. 9–12, 2003.