Vol.2012-MBL-61 No.21
Vol.2012-UBI-33 No.21
2012/3/13

超個人化プロファイル生成のための Web ライフログの分類分析

中 村 明 順[†1]　西 尾 信 彦[†2]

ライフログの注目と Web 利用時間の増加に伴い，ユーザの Web 上での行動履歴を利用したサービスが Web 上では普及している一方で，ユーザが実世界に出かけたときにそれを利用したサービスがないため，ユーザの実世界での活動を支援することが最終的な目標である．ユーザの複数の興味・関心が混在している Web 上の行動履歴を実世界へと適応させるさい，同音異義語によるユーザ間の類似度増加と，ユーザによって同一 Web ページの閲覧意味が異なるということが問題となる．分類の要素にユーザの行動履歴に含まれる閲覧時刻とリファラーとを利用してクラスタリングすることで解決を試みる．そこで本研究では，ユーザの複数の興味・関心が混在している Web 上の行動履歴を，事前に分類の基準を策定せず分類し，より適応範囲の広い超個人特化プロファイルを生成するための分析を行なう．各要素の比率を決めるためさまざまな距離関数を作成し，クラスタリング結果を分析した．その結果，閲覧時刻とリファラーに関しては均等にするほうが適していることがわかった．

The Classification Analysis of Web-Life-Log for Preparing Super Personal Profile

Akinori Nakamura[†1] and Nobuhiko Nishio[†2]

Rapid growth of study on life-log and increasing time of web browsing, browsing logs have been remarkable source to represent user's activity on the Internet. While services which applies browsing logs are popular on Web, no services reflect these logs to user's activity in the real-world. Therefore, the final purpose of my work is to support an active for a user on real-world. We are confronted by two difficulties. The first is an increasing of the number of similarity between users by homonyms. The second is a difference of a browsing motive for a same page. We attempt to solve these problems by clustering approach using browsing time and referrer which are contained in browsing history. This paper is intended as an investigation of the logs classification analysis for a preparing flexible super personal profile unless we define categories in advance. We prepared a variety metric method to determine the ratio of each element. It was found that the ratio of elements of time and referrer should be equalized.

1. はじめに

自身の行動をデジタルデータとして記録するライフログが注目されている．ライフログには，ブログや Twitter といった文章だけでなく，写真やビデオ，さらには GPS で取得した値といったものがある．こういったライフログを蓄積，解析し，ユーザの活動を支援することを目的とする研究が盛んになっている[1),2)]．我々は，GPS デバイスや加速度センサなどが搭載されている高機能端末を利用して，実世界での行動履歴を永続的に蓄積し，解析している[3),4)]．

一方で，ユーザの Web に対する用途が，検索，コミュニケーション，ニュース，購買といったように多様化しているため，ユーザの Web を利用する時間が増加してきている．また，Amazon をはじめとしユーザの Web 上での行動履歴を利用したサービスが普及しており，そういったサービスの利用によって閲覧する Web ページの数が増えている．さらに，ユーザの Web 上での行動履歴から Web 上での活動を支援する研究も盛んになっており[5)–11)]，ユーザが Web を利用する時間はますます増えている．そのため，ユーザの Web 上での活動時間は，生活の半分を占めているといっても過言ではない．

ユーザの Web 上での行動履歴を利用したサービスが Web 上では普及している一方で，ユーザが実世界に出かけたときにそれを利用するサービスはない．そのため，実世界での活動の幅が広がらず，ますます出かけないという現象が発生している．Web 上での行動履歴はユーザの興味・関心といった特徴（ユーザプロファイル）を含んでおり，Web 上での活動時間が増加しているためその特徴の利用価値が増大しているにも関わらず，Web 上の行動履歴から抽出したユーザプロファイルを，実世界で活動するユーザに反映するサービスがないのは問題である．

そこで，ユーザの Web 上の行動履歴からユーザプロファイルを抽出し，実世界に存在する店舗や商品といったコンテンツとマッチメイキングし，実世界でのユーザの活動の幅を広

†1 立命館大学大学院テクノロジー・マネジメント研究科
Graduate School of Technology Management, Ritsumeikan University
†2 立命館大学情報理工学部
College of Information Science and Engineering, Ritsumeikan University

Vol.2012-MBL-61 No.21
Vol.2012-UBI-33 No.21
2012/3/13

げることを最終的な目的とする．そのために，まず Web 上での行動を履歴として蓄積し，ユーザの複数の興味・関心が混在しているそれからその履歴の特徴を抽出しユーザプロファイルとする．そして，Web 上から実世界に存在するコンテンツを抽出し，ユーザプロファイルと抽出したコンテンツとをマッチメイキングさせる．ここで Web 上での行動履歴のことを Web ライフログと定義し，我々はユーザが閲覧した Web ページを蓄積している．また本研究では，ユーザの複数の興味・関心が混在している Web ライフログを，事前に分類の基準を策定せず分類し，より適応範囲の広い超個人特化プロファイルを生成することが目的である．

本論文は全 6 章で構成している．第 2 章で Web ライフログからユーザプロファイルを生成し，実世界で適用するさいの問題を述べる．つぎに，第 3 章で履歴を用いる研究と文書分類に関する研究を紹介する．第 4 章で Web ライフログの分類に利用する要素を説明し，それらの分析を行ない，距離関数を定義する．定義した距離関数に基づいてクラスタリングしその結果を分析するのが第 5 章である．最後に第 6 章で本論文をまとめる．

2. 問題意識

ユーザの実世界での活動は地理的な制約を受けるため，従来のコンテンツを基にした推薦手法では初めて訪れるような不案内なエリアにおいて推薦を受けられないという問題がある．これに対して，我々はユーザプロファイルをコンテンツではなくコンテンツに対する意見であるコメントを分析し，抽象化する手法を提案している[12]．この研究によって，既知/不案内なエリアに関係なく安定した数のコンテンツ推薦が可能，および既知/不案内なエリアともに，従来の推薦手法と同程度の推薦精度を示すこととがわかった．

そこでまず，本研究でのユーザプロファイルを，先の抽象化手法がそのまま適用可能かを検証する．閲覧した Web ページに付与している HTML の body タグ内の文章をコメントと見なし，これを分析し本研究でのユーザプロファイルを生成する．

その結果，次の二つの問題が発生した．ひとつは，同音異義語による 2 ユーザ間の類似度が増加したことである．同音異義語による問題とは,「オブジェクト」という単語であっても，Java といったプログラミング言語で使われたのか，飲食店内に存在しているような小物を指して使用したのか区別できないことである．これによって，本来似ていないユーザ間の類似度が増加し，似ていると判断されてしまう．

二つ目の問題は，同一の Web ページであっても，ユーザによってそのページを閲覧する理由が異なることである．Web ページには飲食，ニュース，プログラミングに関するものといったように多種多様なページが存在している．そのため，ある飲食店の新店舗開業を知らせる Web ページであっても，ユーザによっては飲食に関する情報であるから閲覧した，あるいは住まいの近辺に関する情報であるから閲覧したといったように，同じ Web ページでもユーザによって意味することは異なっている．

そのため本研究は，これらの問題意識を解決するために Web ライフログを分類する．そして分類結果の Web ライフログ群からそれぞれユーザプロファイルを生成し，それをコンテンツとマッチメイキングさせることによって，実世界でのユーザの活動を支援する．

3. 関連研究

関連研究にはユーザ個人の閲覧履歴を利用する研究と，複数ユーザからの履歴を用いる研究，さらに文書を分類する研究とがある．個人の履歴と文書分類に関しては次節以降で詳細に述べる．山元ら[13] は，Web サイトの複数ユーザにおけるアクセスログを基に Web ページを推薦している．これによる推薦は各サイト内の Web ページに限定されるため，適用範囲は狭い．安川ら[14] は，複数のユーザが検索したさいのクエリのログからクエリの関連語を抽出し，これらをクラスタリングする手法を提案している．クラスタリング結果から関連語のクラスタに Web ページを対応づけて提示することで，ユーザの Web 上の行動を支援している．

3.1 閲覧履歴からユーザプロファイルを抽出し利用する研究

松尾ら[5] は，ユーザ個人が閲覧した Web ページの履歴を基に，そのユーザにとって重要である語をハイライトすることで，ブラウジングを支援するシステムを構築している．ユーザが閲覧したすべての Web ページから頻出語を抽出し，これら語の集合との共起の偏りが大きい語をそのユーザにとって重要な語とする．共起の偏りが大きいということは，Web ページの作成者が意味的なつながりを考慮して記述したのであり，その語は Web ページ中において何らかの重要な意味を担っているという考えである．30 分の利用によってブラウジングのしやすさや興味が反映されているかを評価した結果，他のシステムと比較し提案システムは，高い評価を得た．しかしユーザの興味は複数あり，長期的な利用ではこれらが強い影響を与えると考えるが，これに対して分析していない．

杉山ら[6] は，ユーザに関わらず同一検索語には同一の結果を提示するという問題に対して，ユーザが閲覧している Web ページが変わるたびに，その閲覧履歴からユーザプロファイルを更新することで，各ユーザに応じた検索結果を提示している．元来の協調フィルタリングでのユーザ－項目評価値行列を参考にし，ユーザ－単語の重要度行列と見なしてユーザ

Vol.2012-MBL-61 No.21
Vol.2012-UBI-33 No.21
2012/3/13

プロファイルを構築している．しかし Web ページを分類することなくユーザプロファイルに単語の重要度を利用しているため，本研究の問題意識である同音異義語に対して考慮できていない．

河合ら[7),8)] は自動でニュース記事を収集し，ユーザの Web ページの閲覧履歴に基づきニュース記事を動的に分類し，提示するシステムを提案している．収集した記事を分類するために，閲覧した時刻からの経過時間や閲覧回数に基づいて単語の重要度を算出し，さらにその値に基づいてユーザが興味をもっている単語を抽出する．このシステムは，ユーザがすでに分類体系を把握している Web サイトのレイアウトに，収集し分類した Web ページを配置していくことで，大量の情報を効率的にブラウンジング可能にする．さらに，明るい/暗い，承認/拒否といった Web ページに対する印象の尺度を導入しユーザがより共感しやすいように改良している．

大槻ら[9),10)] はニュース記事の閲覧履歴に基づきユーザプロファイルを抽出し，それに適応したニュース配信サービスを提案している．ユーザの複数の嗜好に適応するために，閲覧履歴のみならず非閲覧履歴をも利用して嗜好クラスタを作成している．嗜好クラスタは記事の内容に応じて分類し，閲覧時期に基づいてそれらクラスタに対して重み付けする．しかし，ニュースの配信という限定された用途を対象としている．

3.2 文書を分類する研究

Web ページを含む文書を分類する研究には大きく単語による分類，リンク構造による分類の 2 種類がある．

単語による分類には，たとえば，検索クエリによる結果のうちユーザの求める情報が提示されないという問題に対し，検索結果のページ群を内容に応じて分類する研究[15)] があり，分類に利用するのに適した単語を抽出するために，ファジィ推論を用いている．ファジィ推論とはあるルールに基づいて物事を推論するものである．ほかには，時間が経過するにつれ作成される時系列のニュース記事群において，続報・派生記事の出ていない単独の記事をフィルタリングすることで，クラスタリングの精度を向上する研究[16)] がある．単独記事の検出には固有名詞情報や地理的情報を用いている．さらに，複数の話題で共有される単語やクラスタに誤配置された少数の文書などからの影響を受けないように，各文書間の共通性を分析しそれに基づいたクラスタリング手法の提案[17)] がある．この研究は各文書の文を単位とした単語の共起性を用いて，文書間類似度を算出しクラスタリングしている．Web ページの内容ではなく，その形式である HTML タグに着目した分類手法[18)] がある．しかしこれらの研究は，ユーザに関係なく分類しているため常に同じ分類結果となる．

リンク構造による分類には，たとえば，Web ページ間の内容が類似している場合，リンク構造も密になると考え，この有向グラフに対して最大流アルゴリズムを用いて Web ページをクラスタリングしている[19)]．あるページのリンクは参照元のページと何かしらの関係があるという仮定に基づいている．ほかにはハイパーリンクが共起している Web ページ間は内容が似ているという考えに基づいて，ハイパーリンクからクラスタリングした結果と，内容の類似度からクラスタリングした結果とを重ね合わせる手法[20)] がある．またポータルサイトでのディレクトリ構造を作成するために，ハイパーリンク構造とアンカーテキストに基づいて Web ページを分類している[21)]．単語による分類と同様に，常に同じ分類となりユーザに特化していない．

4. Web ライフログの分類分析

Web ページを自動的に分類する方法として，主にカテゴライズとクラスタリングが注目されてきた[22),23)]．カテゴライズは事前に分類の基準となるカテゴリを設定し，各カテゴリの正解となる教師からカテゴリごとの特徴を抽出し，それを元に未分類な Web ページを分類するという手法である．一方のクラスタリングは教師を利用せず，与えられた未分類の Web ページをまとまりのあるクラスタに分類する手法である．事前に正解知識を入手できないような場合において，クラスタリングは特に有効であることがわかっている．

本研究では，ユーザによって閲覧する意味が異なるという問題のため，事前に分類の基準を策定しないクラスタリング手法を採用する．図 1 に示すように，ユーザプロファイルを各クラスタから抽出することで，あるクラスタで利用されている単語と別のクラスタでのそれを区別できる，つまり同音異義語問題に対処できると考える．また，個人に特化したクラスタリング結果となるように考慮することで，閲覧意味の異なり問題に対応できると考える．クラスタリングには，Web ページ間を評価するための距離関数と，距離関数で利用する Web ページがもつ素性とを決定する必要がある．そこで，次節以降で素性となる要素を分析し，距離関数を定義する．

4.1 クラスタリングに利用する要素

分類で一般的に利用しているのは対象となる文書の内容で，その類似度に基づいてクラスタを形成していく．さまざまな分析研究によって，内容による分類は一定程度の成果があることがわかっている．本研究では分類する対象はユーザが閲覧した Web ページで，その内容はページに記述されている文章である．そこで，この文章を形態素解析し重み付けしたものを，単語の特徴ベクトルと呼び，Web ページの内容を表しているものとする．そし

Vol.2012-MBL-61 No.21
Vol.2012-UBI-33 No.21
2012/3/13

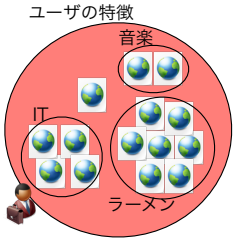

図 1 ユーザプロファイル抽出方法

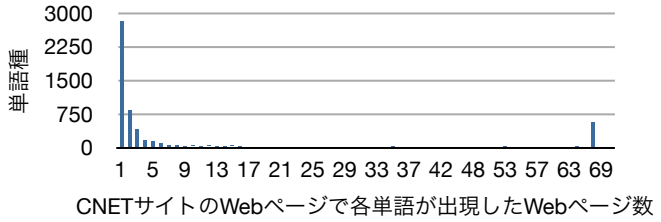

図 2 CNET サイトにおける単語の出現ページ数とその種類の関係

て，Web ページ間の内容が似ているならその距離を近くする．しかし，単語の特徴ベクトルによるクラスタリングは，Web ページ間の静的な類似度算出のため個人に特化していない．このため，次の 2 つの要素によって Web ページのクラスタリングを個人に特化させる．

個人特化のために利用する一つ目の要素は，ユーザが Web ページを閲覧した時刻である．Web ページ間の閲覧時刻が近ければその距離を近くする．なぜなら，ユーザはある活動に関する Web ページをまとめて閲覧するという性質を持つためである．たとえば，ユーザが住まいの近辺に関する情報を調べているときに，ある飲食店の新店舗開業を知らせる Web ページを閲覧すると，これは飲食ではなく，住まいの近辺に関する情報であるから閲覧したと考えられる．閲覧した時刻が近いほど Web ページ間の距離を近くすることによって，問題意識の二つ目に対処する．閲覧時刻によって，各 Web ページにおける単語の特徴ベクトルを重み付ける方法があるが，単語が出現していなければ Web ページ間の距離は近くならず，またユーザによる閲覧の意味の異なりに対応していないため本研究では利用しないこととする．

二つ目は Web ページのリンク元 URL であるリファラーを用いて，クラスタリングの個人特化を目指す．Web ページには Web ページ間の意味的な結びつきを明示的な構造で表しているハイパーリンクがある．ハイパーリンク構造を頂点が Web ページ，辺がハイパーリンクとし有向グラフで表現し，Web ページ間の距離を算出できる．リファラーはこれらハイパーリンクのうちユーザが選択したリンクであるため，頂点をリファラー，辺がログとした有向グラフによって，Web ページ間の距離を算出する．リファラーによって，Web ページ間の一方が日本語，他方が英語で記述されており，単語の特徴ベクトルでは Web ページ間の距離が算出できない場合でも，距離の算出が可能となる．

4.2 要 素 分 析

我々は，Web ライフログ（単にログともいう）を蓄積しており，ログには閲覧した URL とその時刻，リンク元 URL であるリファラーさらに個人を区別するものを含んでいる．また，クラスタリング手法によって分類する対象は Web ページとし，その定義は URL と単語の特徴ベクトル，自身の URL と同一の URL を持つログ群とする．

本研究で分析する対象は，あるユーザがある期間に閲覧した 3000 件のログ内に存在する 1656 種の Web ページである．その期間は 2010 年 6 月 29 日から 8 月 5 日で，おおよそ 1ヵ月である．

4.2.1 単語の特徴ベクトル

単語の特徴ベクトルの重み付けに TF・IDF 手法[24] を用いて，距離算出方法をベクトル間のコサイン距離としクラスタリングすると Web サイトごとにクラスタが生成された．これは各 Web サイトが独自で用意しているレイアウトのテンプレートで利用している単語によって，各 Web サイトの Web ページ間の内容が似ていると判断されたと考える．これを分析するのに，Web ライフログに多く出現した CNET サイトのログを用いる．CNET サイトの Web ページのうち，各単語が出現した Web ページ数を算出し，その値を横軸に，縦軸にその値となった単語の種類としたものを図 2 に示す．CNET サイトの Web ページは最大で 71 となっており，そのうち 68 の Web ページにおいて 565 単語が出現している．したがってこれらの単語によって，Web サイトごとの Web ページがまとまった．

そのため，ある Web ページ i における単語 $t_{i,k}$ を式 1 に示すように重み付けることとする．距離算出方法はコサイン距離のままである．

Vol.2012-MBL-61 No.21
Vol.2012-UBI-33 No.21
2012/3/13

$$t_{i,k} = TF_{i,k} * IDF_{全,k} * IDF_{ド,k} \quad (1)$$
$$IDF_{全,k} = \log(N_{全}) - \log(DF_{全,k}) \quad (2)$$
$$IDF_{ド,k} = \log(N_{ド}) - \log(DF_{ド,k}) \quad (3)$$

$TF_{i,k}$ は Web ページ i 中に単語 k が出現した回数，$IDF_{全,k}$ は全 Web ページ中での単語 k の非出現頻度，$IDF_{ド,k}$ は Web ページ i のサイトのドメインを含む全ドメイン中での単語 k の非出現頻度をそれぞれ示す．なお，$DF_{x,k}$ は x における単語 k を含む Web ページの個数，N_x は x における Web ページの個数をそれぞれ意味し，x には Web ライフログの全 Web ページか Web ページ i のサイトのドメインでの Web ページとなる．ドメインにおける単語の非出現頻度を積算することで，そのドメインにのみ頻出する単語の重要度を下げ，Web ページが Web サイトごとにまとまることを防ぐ．この結果，ほかのドメインをもつ Web ページとまとまる傾向となった．

4.2.2 閲覧時刻

閲覧時刻による Web ページ間の距離算出のために，ログの発生間隔を分析する．図 3 に閲覧した Web ページの回数を横軸，その値をとった Web ページの種類を縦軸に示す．この結果，全 Web ページのうち 99%は 13 回まで閲覧すること，また 1 回しか閲覧しない Web ページは 74%であることがわかった．最大の 81 回となったのは Google のトップページで，その後 Google Reader が続き，大学，天気予報，e コマースといったポータルサイトのトップページが多く見られる．一方で，一度閲覧した Web ページを再閲覧するまでの間隔を示したのが図 4 である．図 4 は同一 Web ページのログのうち最も古い時刻を基準とし，そこからの再閲覧するまでの時間に基づいて，分散を算出し平行根とした値が，0〜6 なら青に，6〜60 なら緑にとして作成したものである．これから 6000s 以内に同一 Web ページを再閲覧するのは，約 87%であることがわかる．また，閲覧の間隔が極端に離れるのは約 180 分程度である．したがって，ユーザの Web 上での活動時間は 180 分程度であると考えられる．

これら分析結果から，閲覧時刻による Web ページ間の距離算出を図 5 のようにする．まず，i) 閲覧の間隔が 180 分を超えるとログ間の関連はないと見なし，ログを切り分けてまとまりを抽出する．つぎに ii) 各 Web ページのログは少なくともひとつのログとの距離をもち，この距離は抽出したまとまりにおける最短ホップで算出する．最後に，iii) 各ログ間の距離の合算の逆数を 1 から引いて正規化した値を，閲覧時刻による Web ページ間の距離とする．ログの切り分けのさい，ある活動の最中にその活動と全く関係のない Web ページを閲覧し，これがクラスタのノイズとなることが考えられる．しかし，Web ページの内容に

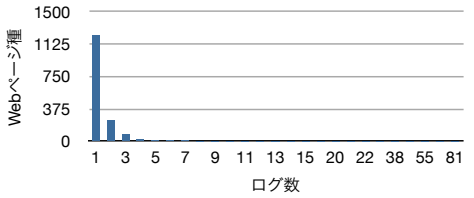

図 3 ログ数と Web ページ種の関係

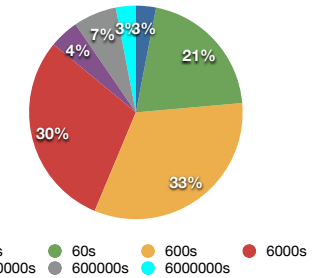

図 4 同一 Web ページの再閲覧までの間隔

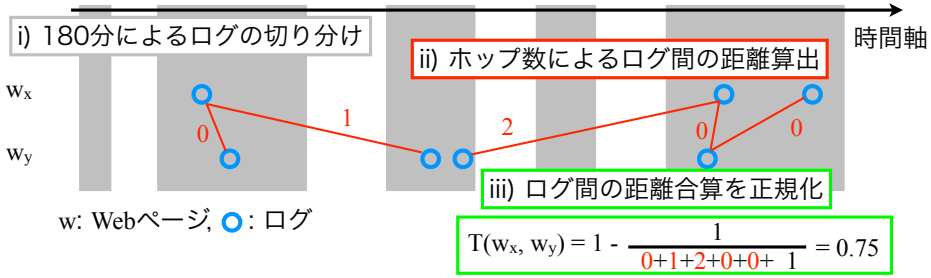

図 5 閲覧時刻による Web ページ間の距離算出方法

関しては単語の特徴ベクトルで考慮しているため，ログの切り分けには時刻のみで行なう．

4.2.3 リファラー

リファラーを頂点，辺にログで，その重みがログの出現回数の逆数とした有向グラフ（本研究では Ref グラフと呼ぶ）と，それを用いた Web ページ間の距離算出方法を図 6 に示す．リファラーによる Web ページ間の距離は，Ref グラフにおいて Web ページ間を結ぶ辺の重みが最小となる経路での値を正規化した値とする．ただし経路を構成できない Web ページ間の距離は最大値の 1 とする．ログの出現回数が多い Web ページ間は，そのユーザにとってログを構成している URL とリファラーとの関連度が高いと考える．

ここで，Google Reader や Google のトップページといったような，本来関係のない Web ページを集めている Web ページの存在が問題となる．さらに，それらをリファラーとして保有している Web ページは全体のうち約 4 割を占めていることがわかった．このため，大

Vol.2012-MBL-61 No.21
Vol.2012-UBI-33 No.21
2012/3/13

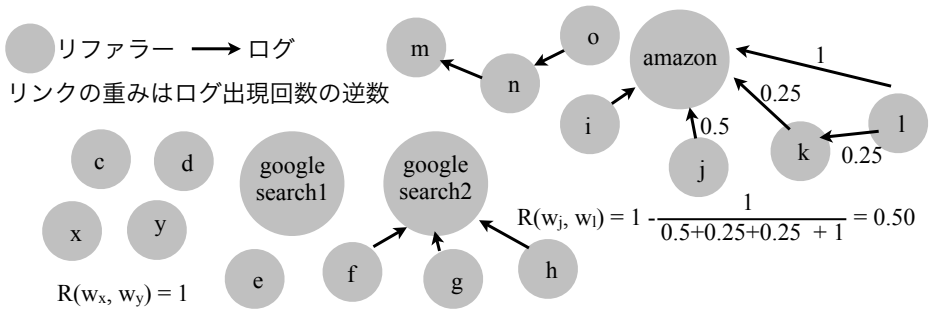

図 6 Ref グラフによる Web ページ間の距離算出方法

量の Web ページから参照を受けている Web ページが，リンク元 Web ページとして機能しているかを判断し，機能していないならその Web ページは Ref グラフに追加しないようにする．具体的には，全 Web ページの 5%以上からのリファラーを有している Web ページのうち，全リンク先 Web ページ間での平均コサイン距離の値が 0.98 以上であれば，その Web ページはリンク元 Web ページとして機能していないと判断する．単語の特徴ベクトルは，ある Web ページがリンク元 Web ページとして機能するかどうかの判断みに利用し，リファラーによる Web ページ間の距離算出には利用しない．

4.3 距離関数

第 4.2 節での分析の結果，Web ページ間の距離関数を式 4 のように定める．

$$dist(w_i, w_j) = \alpha \mathrm{V}(w_i, w_j) + \beta \mathrm{T}(w_i, w_j) + \gamma \mathrm{R}(w_i, w_j) \text{ (ただし} \alpha + \beta + \gamma = 1) \quad (4)$$

$\mathrm{V}(w_i, w_j)$ は単語の特徴ベクトルによる距離で，コサイン距離で算出する．$\mathrm{T}(w_i, w_j)$ は閲覧時刻による距離，$\mathrm{R}(w_i, w_j)$ はリファラーによる距離をそれぞれ表している．これらはすべて 0〜1 で正規化しており，また要素ごとの重要度に基づいた係数を積算し，Web ページ間の距離を算出する．つぎにこの重要度を表す比率を決める．

各要素の比率を決めるため，比率を均等にした場合の Web ページ間の距離と各要素の距離とにおける順位と距離の分布をそれぞれ図 7 から図 10 に示す．各要素の距離を加算した距離の昇順のうち上位 10000 件での各要素の距離を図 8 から示している．図 8 と図 10 から，約 7000 位まではなだらかに距離の値が増えていたのに対し，それ以降は一度距離が 0 に近づいて，またなだらかに増えていったことがわかる．一方図 9 では，約 7000 位までは 0 の距離を多く有し，それ以降距離が 0 になっていない．つまり，上位での閲覧時刻による距離の影響力は高く，単語の特徴ベクトルとリファラーによる距離は二律背反の関係である．

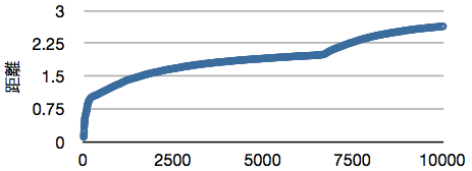

図 7 各要素の距離の加算した距離の分布

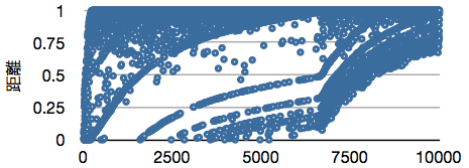

図 8 単語の特徴ベクトルによる距離の分布

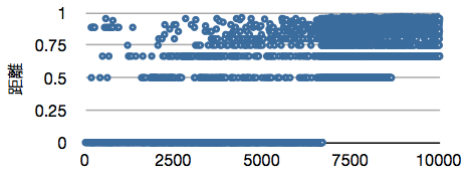

図 9 閲覧時刻による距離の分布

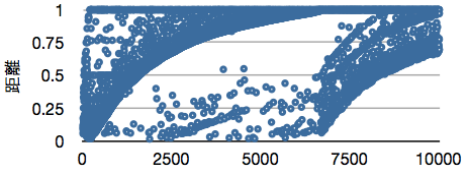

図 10 リファラーによる距離の分布

これら分析結果から，各要素の比率には次の 5 つが考えられる．係数の値が小さいほどそれによる距離が近くなるため，その要素を重視したこととなる．

$$dist(w_i, w_j) = 0.25\mathrm{V}(w_i, w_j) + 0.5\mathrm{T}(w_i, w_j) + 0.25\mathrm{R}(w_i, w_j) \quad (5)$$

$$dist(w_i, w_j) = 0.4\mathrm{V}(w_i, w_j) + 0.2\mathrm{T}(w_i, w_j) + 0.4\mathrm{R}(w_i, w_j) \quad (6)$$

$$dist(w_i, w_j) = 0.2\mathrm{V}(w_i, w_j) + 0.4\mathrm{T}(w_i, w_j) + 0.4\mathrm{R}(w_i, w_j) \quad (7)$$

$$dist(w_i, w_j) = 0.4\mathrm{V}(w_i, w_j) + 0.3\mathrm{T}(w_i, w_j) + 0.3\mathrm{R}(w_i, w_j) \quad (8)$$

$$dist(w_i, w_j) = 0.33\mathrm{V}(w_i, w_j) + 0.33\mathrm{T}(w_i, w_j) + 0.33\mathrm{R}(w_i, w_j) \quad (9)$$

ひとつ目の式 5 は閲覧時刻による距離の上位のほとんどが 0 と係数の値を積算する利点がないため，閲覧時刻の係数つまり β の重要度を低くし，残りを均等にする比率である．式 6 に示すように二つ目の比率は，ひとつ目と対象的に閲覧時刻の重要度を高くし，よりユーザの Web 上での活動を重視することで，Web ページの内容による距離より．Web ページ間の内容の類似度を重要視し，閲覧時刻やリファラーによる距離はそれを補助する比率が三つ目である (式 7)．四つ目の比率が，同音異義語とユーザによって同一 Web ページ閲覧の意味の異なりという問題に対処するために利用している閲覧時刻とリファラーによる距離の重要度を高くし，内容に関してはこれらを補助するために利用する式 8 である．最後にすべての要素を均一に扱うのが式 9 である．5 つの比率によるクラスタリング結果を第 5 章に示し，分析することで 3 つの要素の比率を決定する．

Vol.2012-MBL-61 No.21
Vol.2012-UBI-33 No.21
2012/3/13

図 11　クラスタ内に水出しコーヒーと書籍

図 12　クラスタ内に小麦粉と楕円曲線

5. クラスタリング結果と考察

クラスタ間の距離関数に郡平均法を用いて，第 4.2 節で分析した 3000 件のログ内に存在する 1656 種の Web ページを対象に階層的クラスタリングを行なった．その結果のうち，特徴的なものを図 11 と図 12 とに示す．紙面の都合上途中で切れているのもあるが図には Web ページのタイトルとその URL を記述しており，線で結んでいるのが最短距離となった Web ページの組である．距離やクラスタ内の Web ページの個数に違いがあるものの，図中の Web ページ群はひとつのクラスタ内に存在している．

図 11 は，水出しコーヒーに適している器具や豆に関して調べ最終的に Amazon で器具を購入したときのログと，経営に関する書籍を Amazon や本やタウン[⋆1]に在庫があるかどうかを調べていたときのログのうち Amazon に関するログとがクラスタ内に存在した．この現象は式 5 と式 6 とで起こり，要因は Amazon のトップページというリファラーによって Web ページ間の距離が近くなったからである．これらを調べていた時期の間隔は 20 日程度離れているが，式 6 では時刻による距離の係数を 0.2 と小さくしており，リファラーによる距離が与える影響より強くなったと考える．一方で，単語の特徴ベクトルによる距離は約 0.987 で，これによる影響は閲覧時刻やリファラーによる距離と比べ弱かったため，残りの式 7 や式 8，式 9 では図 11 のようなことが起こらなかった．

図 12 は，第一屋製パンの戦略を論じる講義の課題でパンの原材料である小麦粉の価格変動に関して調べていたときのログと，暗号理論に用いる楕円曲線について調べていたときのログとがクラスタ内に存在するようになった．この現象は式 6 のみで起こり，第一屋製パンや暗号理論に関するクラスタが別に形成されているにもかかわらず，小麦価格と楕円曲線を内包したクラスタが生成された．この要因は閲覧時刻の比率が低く，またこれらの活動の間隔が狭いため，Web ページ間の距離が他の式より著しく近くなったのである．事実，これらは同日の昼夜で調べていたものである．一方で残りの式では，第一屋製パンのクラスタ内に小麦価格に関する Web ページが，暗号理論のクラスタ内に楕円曲線に関する Web ページがそれぞれ構成され，この 2 つのクラスタを包含したより大きなクラスタが作成された．

閲覧時刻を相対的に重視しない式 5，反対に重視する式 6 には顕著な問題が見られた一方で，Web ページの内容を重視する式 7 と，ユーザによって同一 Web ページ閲覧の意味の異なりに対してより重点を当てた式 8 さらにすべての要素の比率を均等にした式 9 に関しては明確な違いを得ることができなかった．内容に関してはそもそもすべての式において一定の成果を得ており，より重視する必要がなかったと考えている．残りの 2 式での 3 要素の比率に大きな差が無く，距離には誤差として吸収されてしまった．言い換えると，多少の階層構造に変化があったが，巨視的に判断すると同一クラスタに内包していたのである．

以上をまとめると，本研究では，各要素の距離の分布から閲覧時刻に対して重要度を高める式と低める式とを作成し，また残り 3 つの式の計 5 つの式に基づいてクラスタリングした．その結果，閲覧時刻のみに対して比率を高めても低めても適切なクラスタが形成されないこと，また Web ページの内容を重視する必要がないことがわかった．閲覧時刻とリファラーの比率を均等に扱う 2 つの式において，内容も均等にするか否かに関しては本研究においては判断がつかなかった．

6. おわりに

本研究は，実世界でのユーザの活動の幅を広げることを最終的な目的とし，ユーザの複数の興味・関心が混在している Web ライフログから，事前に分類の基準を策定せず分類し，より適応範囲の広い超個人特化プロファイルを生成することを目指した．Web ライフログから抽出したユーザプロファイルを実世界へと適応するさい，同音異義語によるユーザ間の類似度増加と，同一の Web ページを閲覧したとしてもユーザによってその行為の意味が異なるという問題があることを指摘した．その対策として，クラスタリングに利用する要

⋆1 現在の Honya Club

Vol.2012-MBL-61 No.21
Vol.2012-UBI-33 No.21
2012/3/13

素に Web ページの内容を表す単語の特徴ベクトルとユーザが Web ページを閲覧した時刻，どの Web ページからの遷移なのかを示すリファラーとを用いた距離関数を定義した．各要素の適切な比率を決めるために，さまざまな距離関数を作成し分析した．その結果，閲覧時刻とリファラーの要素の比率は片方のみを重視せず両者を均等に扱うほうが適していることがわかった．しかし内容に関しては明確な違いが得られなかったため分析できていない．

今後の課題は，実験に用いた期間が 1ヵ月であり，これを 1 年といった長期間にすると閲覧時刻やリファラーの傾向が変化する可能があるため，新たな比率を分析する必要がある．また評価にはクラスタリング結果を分析するという直接評価を行なったが，分類の有無によりその後のサービスにどの程度成果を得られるかを評価することも重要である．ユーザプロファイルを抽出するさい，適切なクラスタの分割方法に関しても分析する必要がある．

参 考 文 献

1) 小柴等, 相原健郎, 森純一郎, 小田朋宏, 松原伸人, 星孝哲, 武田英明. 記憶の想起と記録のためのライフログ・ブログ連携型支援手法の提案. 情報処理学会論文誌, Vol.51, No.1, pp. 61–81, January 2010.
2) 山田直治, 礒田佳徳, 南正輝, 森川博之. 屋外行動支援のための gps 搭載携帯電話を用いた移動経路の逐次的精錬手法. 情報処理学会論文誌, Vol.52, No.6, pp. 1951–1967, 2011.
3) 坂本憲昭, 坂本一樹, 名生貴昭, 市川昌宏, 新井イスマイル, 西尾信彦. 同期シナリオを用いてセンシング携帯端末と協調連携するアプリケーションフレームワークの提案. マルチメディア，分散，協調とモバイル (DICOMO2010) シンポジウム論文集, pp. 591–601, 2010.
4) 藤井陽光, 川崎万莉, AnhTuan Nguyen, 安積卓也, 西尾信彦. 細粒度 wi-fi 測位と加速度センサを併用した屋内行動認識. マルチメディア，分散，協調とモバイル (DICOMO2011) シンポジウム論文集, pp. 720–728, 2011.
5) 松尾豊, 福田隼人, 石塚満. ユーザ個人の閲覧履歴からのキーワード抽出によるブラウジング支援. 人工知能学会論文誌, Vol.18, No.4, pp. 203–211, 2003.
6) 杉山一成, 波多野賢治, 吉川正俊, 植村俊亮. ユーザからの負担なく構築したプロファイルに基づく適応的 web 情報検索. 電子情報通信学会論文誌 D-I, Vol. J87-D-I, No.11, pp. 975–990, 2004.
7) 河合由起子, 宮上大輔, 田中克己. 個人の選好に基づく複数ニュースサイトの記事収集・閲覧システム. 情報処理学会論文誌：データベース, Vol.46, No. SIG 8(TOD 26), pp. 14–25, 2005.
8) 河合由起子, 熊本忠彦, 田中克己. 印象と興味に基づくユーザ選好のモデル化手法の提案とニュースサイトへの応用. 日本知能情報ファジィ学会誌, Vol.18, No.2, pp. 173–183, 2006.
9) 大槻一博, 服部元, 星野春男, 松本一則, 菅谷史昭. 携帯向けオンラインニュース配信のための視聴/非視聴履歴に基づく嗜好クラスタ管理手法. 日本データベース学会 Letters, Vol.6, No.1, pp. 37–40, 2007.
10) 大槻一博, 服部元, 松本一則, 滝嶋康弘, 菅谷史昭, 鹿喰善明. パーソナル・オンラインニュース配信システムの実証実験. 日本データベース学会論文誌, Vol.7, No.1, pp. 43–48, 2008.
11) 堀幸雄, 今井慈朗, 中山堯. ユーザの web 閲覧履歴を用いた検索支援システム. 情報知識学会誌, Vol.17, No.2, pp. 95–100, 2007.
12) 中村明順, 通山和裕, 新井イスマイル, 西尾信彦. 実世界嗜好推薦の coverage 拡大のためのユーザプロファイル抽象化手法. 情報処理学会論文誌, Vol.51, No.12, pp. 2343–2353, December 2010.
13) 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, Vol.48, No. SIG 11(TOD 34), pp. 38–48, 2007.
14) 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.
15) 城市広大, 三好力. ベクトル空間法とファジィ推論を用いた web 検索結果自動分類システム. 日本知能情報ファジィ学会誌, Vol.18, No.2, pp. 184–195, 2007.
16) 中村智浩, 平野孝佳, 平手勇宇. 単独記事フィルタリングを用いた時系列ニュース記事分類法の提案. 日本データベース学会論文誌, Vol.7, No.2, pp. 7–12, 2008.
17) 川谷隆彦. 多文書間の共通性分析に基づく文書クラスタリング. 情報処理学会論文誌, Vol.47, No.6, pp. 1903–1917, June 2006.
18) 折原大, 内海彰. Html タグを用いた web ページのクラスタリング手法. 情報処理学会論文誌, Vol.49, No.8, pp. 2910–2921, August 2008.
19) 大野成義, 渡辺匡, 片山薫, 石川博, 太田学. Max flow アルゴリズムを用いた web ページのクラスタリング方法とその評価. 情報処理学会論文誌：データベース, Vol.47, No. SIG 4(TOD 29), pp. 65–75, March 2006.
20) 高橋功, 三浦孝夫. ハイパーリンクの共起性を用いたクラスタリング手法. *DEWS2005, 1C-i12*, 2005.
21) 鈴木祐介, 松原茂樹, 吉川正俊. アンカーテキストとハイパーリンクに基づく web 文書の階層的分類. 人工知能学会第 19 回全国大会論文集 3C2-02, 2005.
22) Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, Vol.34, No.1, pp. 1–47, March 2002.
23) A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: A review. *ACM Computing Survers*, Vol.31, No.3, pp. 264–323, September 1999.
24) G.Salton. Developments in automatic text retrieval. *Science(Washington, D. C.)*, Vol. 253, No. 5023, pp. 974–980, 1991.