

直近の取得 Web 情報に基づくユーザの現在の興味・指向の予測に関する研究

Predicting User Interests and Orientations Based on their latest Web Acquisitions

土屋 健/Takeshi TSUCHIYA¹・山田 哲靖/Tetsuyasu YAMADA²・広瀬 啓雄/Hiroo HIROSE²・吉永 浩和/Hirokazu YOSHINAGA³・澤野 弘明/Hiroaki SAWANO⁴・小柳 恵一/Keiichi KOYANAGI⁵¹公立諏訪東京理科大学 工学部 准教授・²公立諏訪東京理科大学 工学部 教授³ログリー株式会社 代表取締役社長・⁴愛知工業大学 情報科学部 准教授⁵早稲田大学 理工学術院 教授

[Abstract]

This paper discusses and proposes a method to predict information based on user web behaviors. Conventionally, some of the information stored in cookies is used for recommendations regardless of user presence. This paper develops a method to extract information features associated with user interests using neural networks, and predict the information users are interested in. The method also enables a flexible information service to be constructed using predictions adapted to the user's presence. The evaluations indicate that the method is effective and practical.

[キーワード]

ユーザ興味情報, 文脈の類推, ニューラルネットワーク, コンテンツ特性分析

1. はじめに

インターネットは、ユーザの利用、ネットワーク環境に適応できる柔軟性のあるサービスとして我々の生活インフラとしてその適用領域を拡大し続けている。このとき、最も重要な収益モデルである Web 広告では、その双方向性を活用して、ユーザの投入した検索クエリ、表示しているコンテンツ、過去の web 履歴に適応させた広告を配信する興味関心連動型広告、登録した性別、居住地、年齢などの属性に適応した広告を配信する行動ターゲティング広告など多くの手法が利用されている。これら Web 広告に用いられる手法は、従来のマス広告の市場が急速に縮小しているのに対し、対象ユーザのターゲティングによる情報の到達性、表示されたユーザからの応答性に優れることから、これら機能を提供する配信プラットフォームの利用は増大しており Web 広告の市場はその規模を拡大し続けている [1]。そのため、今後ますます Web 広告配信の機能は高度化が期待され、利用ユーザに適応した配信技術が求められる。

現在のこれら手法による Web 広告の選択手法は、現在表示されているコンテンツの内容、ユーザが入力した検索クエリに含まれる単語、キーワードを直接的に含むもの、つまり利用ユーザが明確に認識する単語、キーワードを含む広告を配信する。もしくは、利用ユーザが過去に訪れた履歴のある Web サイトからの広告が配信されるいずれかのケースが一般的である。つまり、利用ユーザの既知である単語、キーワード以外を含む関連する情報、訪問したことのない Web サイトの情報といったユーザにとって未知の情報は提供を前提としていない。従って、利用ユーザの興味ある情報を提供しているとは言い難く、Web 広告の配信では今後早急に求められる技術の一つであると考えられる。加えて、利用ユーザの興味情報が分析可能となれば、Web 広告などの情報推薦として利用だけでなく、Web 負荷分散のためのキャッシュサーバへの適用、クラウド環境での分散データ管理への適用といったシステム・ネットワーク技術への応用による情報インフラとしてインターネットの高度化が期待できる。

利用ユーザの興味は、大きく分類すると個人の趣味や志向など時期や期間に関わらず変わりなく長期に渡って継続する興味と、直近の課題や、調査など必要性によって発生する一定期間に集中する短期間の興味があると考えられる。そこで筆者らは、利用ユーザが取得する情報に着目し、これら情報の特徴を解析することにより利用ユーザの興味を予測する手法を検討する。本研究では、ユーザの興味の遷移が半断しやすいため短期的なユーザの興味

の予測のために、ユーザが直近に取得している情報に着目する。本手法の利用により、従来の Web マーケティング、Web 広告の高度化、ユーザの行動予測に基づいた情報システムの動的な設計などの応用を期待できる。

1.1. 仮説

本研究では、ユーザの短期の興味は最長で直近 30 分の継続が限度であると仮定し、この範囲内では各ユーザの取得する web コンテンツ情報にはそれぞれの現在の興味情報が含まれていると考えることができる。従って、取得したこれら Web 情報の特徴解析により、ユーザの興味情報を予測できると仮説を立てる。

各ユーザ興味の特徴解析には、隠れ層を 1 つのみ有する一般的なニューラルネットワークにおいて教師あり学習を行うことにより、ユーザ取得情報の回帰によるユーザ興味予測が可能となる。このとき、学習に利用する教師データは、Web 広告をクリックしたユーザの Web 上の行動履歴を利用する。

Web 広告は、前述した手法により、ユーザおよび、コンテンツに適応した広告が選択される。従って、クリックするユーザにとって広告に含まれる情報が少なからずその時点における自身の興味を示す、もしくは類似すると仮定できる。そのため、このクリックしたユーザの Web 上の行動履歴は教師データとして利用が可能であるとする。また、提案手法との比較として、情報に含まれる単語の統計的性質である出現頻度を利用した特徴解析手法を用いた場合と性能比較を行い、提案手法の有効性を示す。

以降、2 章において、対象となる web 行動履歴から分析の対象となるユーザの興味を示す情報の取得手法についてまとめ、3 章で提案手法となるこれら情報の解析手法を明らかにする。4 章で提案手法の評価を行い、5 章に考察、そしてまとめる。

2. ユーザの興味情報の取得

本章では、ユーザの興味情報の予測に関連した研究と、このユーザ興味と各ユーザの web 上の行動履歴の関係について検討する。

2.1. 関連研究

本研究の関連研究として、文献 [2] では、本稿と同様にユーザの Web 履歴の分析に基づいて、長期的・短期的ユーザの潜在的な興味を分析する手法を提案している。特徴分析には、ユーザの訪れたサイトの FQDN、Web サイトのカテゴリを用いた場合の比較を行っており、Web サイトのカテゴリを利用した手法の方が性能が良いこと、分析期間の長短は、予測性能に関連性がないことを明らかにしている。分析に Web コンテンツの内容を用いる本稿の提案手法と比較して、FQDN、Web サイトのカテゴリの特徴の学習による分析は、テキスト処理が発生しないため処理負荷は小さい。しかし、どちらの情報の特徴解析に利用しても Web サイト特徴の分類しか分析できず、ユーザ興味情報の分類のみの予測となる。本稿の提案手法では、処理負荷は大きくなるが、コンテンツに含まれる情報を特徴として解析するため、興味情報の分類より詳細に、コンテンツ内容まで考慮した類似性による判断が可能となる。また、文献 [3] では [2] の手法において、閲覧履歴の順序性を考慮した RNN (Recurrent Neural Network) を分析に導入することで予測性能の向上を期待できることを明らかにしている。この手法では順序性を利用することで、閲覧履歴の前後関係が明らかになり、[2] よりも短期のユーザの興味情報を強調する効果が得られるが、[2] と同様に Web サイトの分類、つまりユーザ興味の分類の予測のみが可能となっている。文献 [4] では、EC サイトの購入前後の行動から、購買行動を予測する手法を提案している。特に、ロジットモデルをベースとしたモデルを提案している。しかし、当該手法ではあくまで統計的予測であるため、利用者の興味まで予測はできず、それまでに発生していない行動パターンを予測することはできない。本稿では、取得したコンテンツの情報自体を解析していることから、コンテンツ間の類似性、相関性の観点からユーザの興味情報を予測することが可能となっている。

2.2. ユーザの興味を含む情報

一般的に、ユーザは各時点において、興味がある/知る必要がある情報に対して、検索/取得に web を利用していると考えることができる。このとき、その時点で検索/取得する情報には、一定の共通する情報が含まれており、この共通情報がユーザの興味ある情報であると 1.1 節で示していた。しかし、この短期的な興味の継続は、上記の検索/取得により結論が得られた、または時間の経過とともに大きく変化する可能性が高いため、本稿では同一の興味の継続が最長 30 分と仮定していた。これは、文献 [5] において、人間(子供)の集中力の継続は、15 分単位であり、最長で 45 分であると示されていることから、ユーザ興味の継続も同様と考え、本稿では中間となる 30

分を興味継続の限界と設定した。つまり、最大でもこの時間の範囲内が同一の興味が継続する時間であり、web 上の行動履歴の特徴もこの範囲時間内で変化すると考える。図 1 にある時刻($t=0$)における各ユーザの web 行動履歴と、各ユーザの取得する web 情報を示す。図 1 に示すように、設定した同一の 30 分の範囲においても、ユーザによって取得する Web 情報量は異なる。

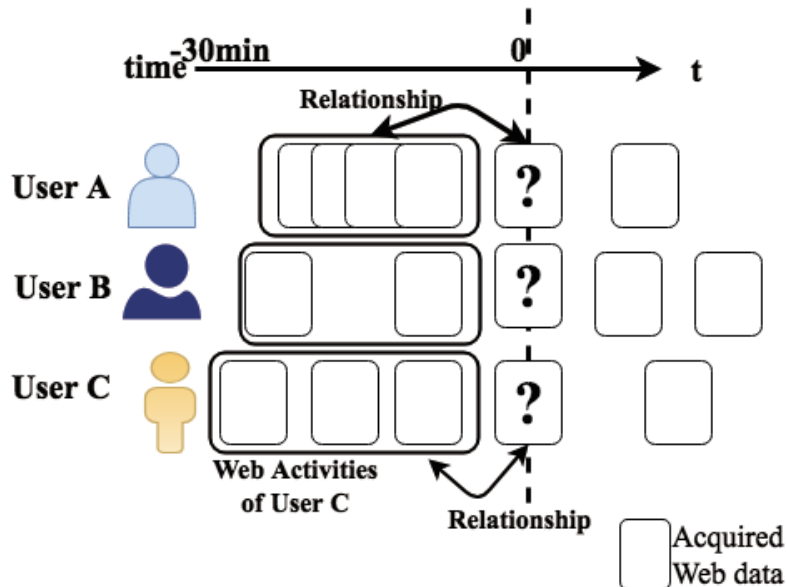


図 1 ユーザが Web で取得する情報

2.3. 取得する Web コンテンツ

ユーザが取得する web コンテンツは、本文に相当する静的コンテンツ単体(本稿ではこの部分をメインコンテンツとする)と、実行時の環境に適応して生成されるコンテンツ(例えば、類似コンテンツへのリンクなど)と組み合わせにより構成される。加えて、レイアウト情報、画像のリンク、Web 広告といった複合した情報から web コンテンツは構成される。つまり、ユーザが求める情報は、静的コンテンツに記述されていると考えられ、メインコンテンツのみが必要となる。しかし、実際の表示時には、付随する情報も同時に取得している。そのため、ユーザの現在興味ある情報を精度高く抽出するためには、メインコンテンツのみを対象として解析が必要となる。

2.4. Web 上での行動履歴

各ユーザの Web 上での行動履歴に相当する取得した Web コンテンツの URL 情報は、ユーザの web コンテンツ取得時に Web 広告の決定のために配信される情報から、ユーザ環境とともに収集される。このとき、本稿では、初アクセス時に生成されるユーザ ID 情報、時刻情報が取得され、データベース上で管理する。

取得した情報の解析には、ユーザ単位で前述した URL のリスト作成し、リストに基づき順次 Web コンテンツをスクレイピングにより取得する。本稿では、WebKit ベースのヘッドレスブラウザの PhantomJS [7]、自動化ツールである Selenium [8] を利用して自動的に巡回することで取得している。このプログラムは、クライアントとして各 URL へアクセスし、コンテンツを取得する。これにより、各ユーザが Web 上で取得した情報を得ることができる。

2.5. メインコンテンツの推定と情報の抽出

取得した Web コンテンツから、メインコンテンツを抽出するには、これまでに、自然言語手法を利用してテキストの内容に基づきメインコンテンツを推定する手法 [9] と、機械学習による web ページ内でのテキストの配置場所の学習によるメインコンテンツを推定する手法 [10] など多くの手法が検討されている。

文献 [11] では、メインコンテンツは最も内容(文字数)が多いと仮定し、実コンテンツによる評価からメインコンテンツは全体のテキストの平均 78.8%を占め、最も大きい割合を有していることを明らかにしている。そこで本稿では、`<body>~</body>`までの全体のテキスト量に対する各ブロックの割合を導出してメインコンテンツ

の箇所を推定する。このとき、文字数のカウント処理のみしか発生しないため、前述した関連研究よりも導出に要する処理負荷は小さくなる。

日本語で記述された Web コンテンツの抽出には、文章の解析と各語への分解が必要となる。このとき、テキスト中では、単語は助詞、助動詞と組み合わせ・活用が行われていることから、英単語のようにそのまま分解できない。従って、単語の抽出の実行には、形態素解析による分析を実行し、テキストを単語と品詞へ分解する。その後、情報の文脈、コンテキストとは無関係と思われる語、助詞、接続詞、数詞といった情報の内容に関わらない語は、処理負荷の低減の観点から削除する。本稿では、オープンソースの MeCab [12] を適用することで、形態素解析を実行し、品詞に分解、単語を抽出している。

本稿では、この導出された情報が図 1 における各ユーザがそれぞれ web 行動において、取得した情報に相当する。そして、本稿では 1.1 節で示したようにこれら情報にユーザの興味情報が含まれると考える。

3. 情報の学習

本章では、各ユーザの取得した情報に対して、教師データとしてニューラルネットワークにおいてその特徴を学習してモデルの構築手法と、各ユーザの取得した情報の特徴解析による回帰手法を示す。図 2 では、対象とする時間の範囲(図中 Target Time に相当)において、学習用のデータ(図中 Model A~C に相当)によるモデル構築と、ユーザが取得した情報に基づく回帰(図中 User 1 に相当)による提案手法の概要を示す。

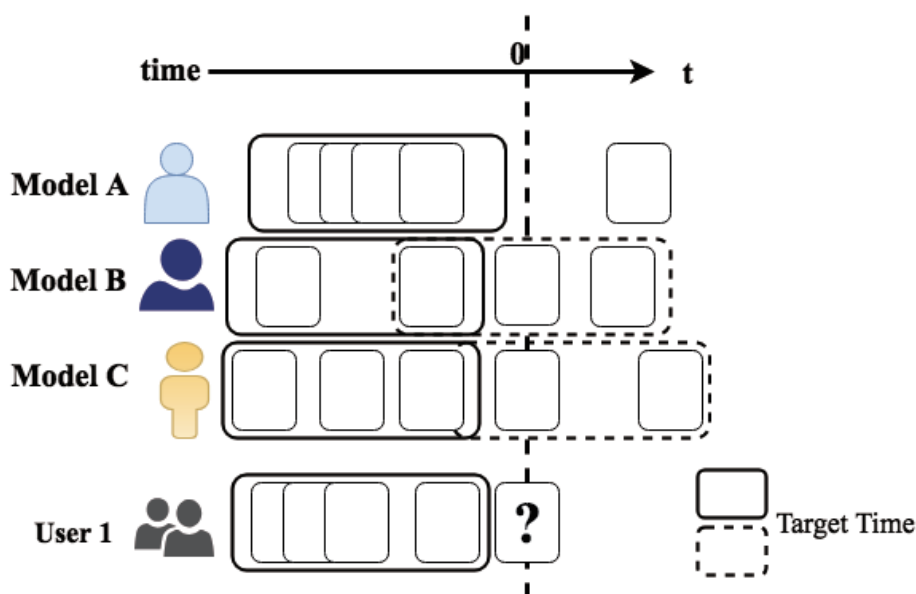


図 2 取得した情報に基づくユーザ興味情報の予測

3.1. 学習に利用するデータ

1.1 節でも示したが、提案手法では、隠れ層を 1 つ持つニューラルネットワークにおいて教師あり学習によるモデル構築と、各ユーザの興味情報に回帰による予測を行う。学習に利用する教師データは、図 2 に示す Web 広告をクリックしたユーザの取得した Web 情報が用いられる。

現在の Web 広告は、Web コンテンツに付加されている広告枠に対して、1 章の冒頭で示したようにこのコンテンツに関連、もしくは類似する属性を持つ広告であり、ダウンロードするユーザの属性(年齢、生活圏など静的な情報、過去の Web 履歴など動的情報)に適応した広告の入札により、配信される広告が動的に決定している [6]。従って、Web コンテンツに表示される Web 広告をクリックするユーザは、現在 Web 広告に興味ある、もしくは誤って Web 広告をクリックしたどちらかであると考えることができる。従って、対象となるデータ数の増加に伴い、誤って Web 広告をクリックするユーザよりも Web 広告に興味があつてクリックしたユーザが増加すると考えることができる。そのため、本稿ではこの Web 広告をクリックしたユーザの Web 上で取得した情報は興味を示してい

と考えることができ、学習用の教師データとして利用する。

3.2. 取得情報の文脈学習

本提案手法では、抽出したユーザの取得した情報の特徴解析には、隠れ層を1つ有するニューラルネットワークにおいて、教師あり学習を行う。このとき、興味が継続していると仮定した時間(図2 Target Time に相当)を対象として教師データを用いたモデルの構築を行う。具体的な手法としては、PV-DM(Paragraph Vector—Distributed Memory)モデル [13] を適用してニューラルネットワーク上で出現単語の語順を学習する。このPV-DMモデルでは、図3左部に示すように、入力パラメータであるwindow size に含まれる単語、出力がウィンドウに隣接する次の語となるようにニューラルネットワークのパラメータ学習を進め、取得した情報に含まれる文脈の学習を行う。

図3では、パラメータ'window size=3'words が設定された環境である。このとき、'target word D'を出力するために'word A' ~ 'word C'を文脈として学習する。その後、順次ウィンドウをずらしていくことで、情報に含まれる語全てに対して入力に対する出力を学習する。図3では対象を右にスライドさせることから、'word E'を予測するために'word B' ~ 'word D'を文脈として学習する。加えて、本稿の提案手法では、逆方向つまり、図3右部入力をウィンドウ内の単語'word E' ~ 'word G'、出力をウィンドウの直前の語'word D'とした学習も行う。つまり、ウィンドウの両側に存在する語を予測するようにニューラルネットワークでの学習を行う。これにより、単語数の少ない情報でも効率的な学習を期待できる。

以上のプロセスから、図2におけるWeb広告をクリックしたユーザの履歴である学習用のデータからモデルが構築される。このモデルの構築により、学習した特徴に基づいた各語のベクトル化と、各ユーザの取得する情報のベクトル化が可能となる。

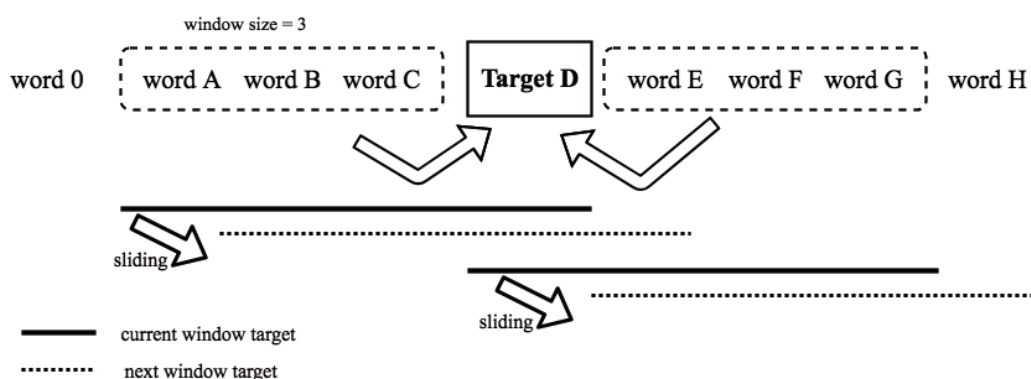


図3 文脈の学習

3.3. 興味情報の予測

ユーザ興味情報の予測を行うためには、学習用のデータと同様に、予測を行う各ユーザが直近までに取得した情報を構築したモデルを利用して、ベクトル化を行う。このとき、ベクトル化された各ユーザの取得情報は、ユーザの特徴を示している。(図2のUser1の"?"の部分に相当)。そのため、予測される情報は、この特徴を示すベクトルへ回帰する情報が選択されることになる。しかし、ベクトル情報のままでは予測として利用性が低い。そのため、あらかじめ予測候補となる情報(例えば、予測候補となるWeb広告やWebコンテンツ)をベクトル化して準備しておく。これにより、予測候補である情報とユーザの興味情報のベクトルとの類似性を導出し、最もユーザ興味情報に近いものを予測情報とする。つまり、2ベクトルのなす角を計測するcos類似度によって、最も類似する情報を判定することになる。

4. 評価

本章では、実Webサービスから取得されるユーザのWeb上での行動履歴から、取得した情報を抽出することで提案手法の評価を行う。

4.1. 評価に利用する Web 行動履歴

本評価に用いるユーザの Web 行動履歴は、学習用の教師データ、評価用データどちらも 3.1 節で示した Web 広告をクリックしたデータが利用される。本稿で利用されるこのデータは、ある事業者が提供する Web コンテンツ内にレコメンド、広告を表示する広告枠(媒体)と、この広告枠へ出稿する広告主が接続されるアドネットワークへ通知される Web 広告の利用ログを利用する。このとき、対象となる媒体は、特定のドメインでなく、様々なドメイン上の不特定の Web サイトに配置されている。

ユーザ ID は初回の広告枠取得時にランダムに生成され、同じアドネットワークに接続する広告枠の取得時には同一のユーザ ID が利用される。従って、ユーザ ID の追跡によるユーザの Web 上での行動履歴を取得が可能となる。

本評価では、表 1 に示す学習用に約 5000 件の web 上での行動履歴(ユーザ数は約 1300 ユーザ)でモデルを構築し、150 ユーザの行動履歴を評価として予測を行う。

4.2. 評価方法

本評価では、表 2 の開発環境を用いて、表 1 の Web 行動履歴データを教師データとして、学習させてユーザの興味情報の予測モデルを構築する。そして、評価用データによるユーザの興味情報を予測する。予測した情報の性能評価は、4.1 節で示したように評価用データも Web 広告をクリックしたデータであることから、クリックした広告を正解データとしてみなすことができ、予測とこの正解データとの類似性に基づいて精度判定を行う。

比較として、同一の学習データを用いて、情報に含まれる単語の統計値である語の出現頻度を用いた従来型手法と比較する。この従来手法は、情報検索に用いる手法であり、学習用のデータを利用してモデルを構築する。このとき、情報の特徴に含まれる単語の出現頻度で表す $tf \cdot idf$ (Term Frequency · Inverse Document Frequency) [14] を利用する。さらに、構築モデルの特徴を強調するために LDA (Latent Dirichlet Allocation) [15] による確率的次元削減を適用する。

表 1 評価用ユーザの行動履歴

Environment	Specification
Web 行動履歴データ件数	4500
ユーザ数	1300
評価用データ数	150 ユーザ

表 2 評価環境

Environment	Specification
前処理	gensim [16]
モデル構築	Tensor Flow 1.8 [17], Python 3.6.2
OS	Ubuntu 16.10, Docker 17.09

4.3. 評価結果

提案手法の評価結果を図 4 に示す。x 軸は適合率(Precision rate), y 軸は再現率(Recall rate)を示している。図中では、提案手法“Proposal”, 統計的手法を利用した“Conventional”それぞれを示している。一般的に、適合率-再現率グラフでは、右上側に表されるほど性能が優れていると判断され、この観点では、提案手法は情報の特徴を表現できていると判断できる。

しかし、提案手法は図中で性能が安定していないことが観測できる。これは、現在のスクレイピングによる Web コンテンツ収集を利用する場合では、ページ内のメインコンテンツに加えて、動的に生成されるヘッダ、サイドの広告、SEO(Search Engine Optimization)用のタグなどが付加されている。これら情報は、現在のメインコンテンツに類似する情報、cookie などユーザの Web 履歴に依存した情報が動的に付加されるため、これら内容がメインコンテンツの情報を強調するケース、影響しないケースにより、性能が区間により揺らぐことで安定しないと考える。しかし、前述したように全区間において統計的性質を利用した従来手法による予測よりも、教師あり学習を用いたニューラルネットワークによる提案手法の方が全区間における性能は優れていると判断できる。

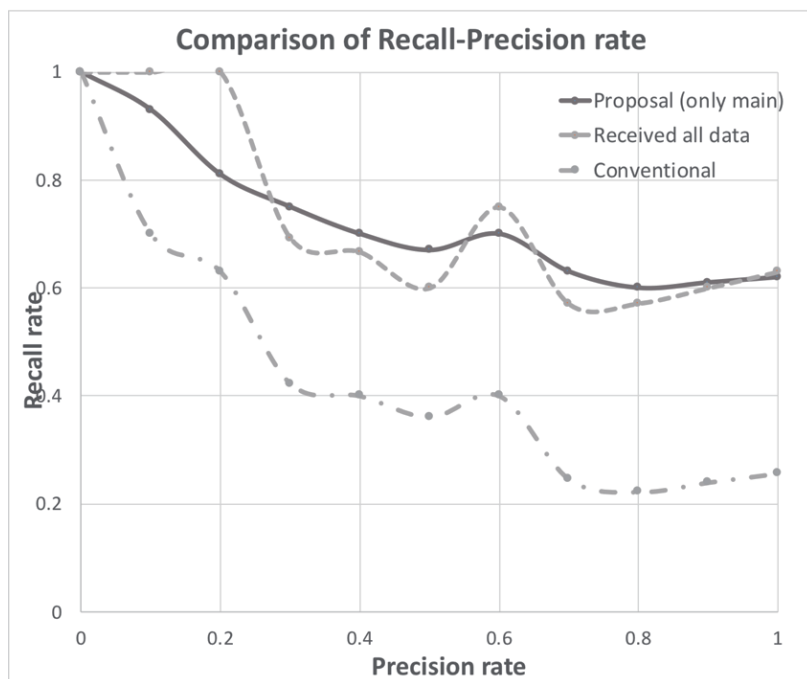


図 4 精度-再現率による比較評価

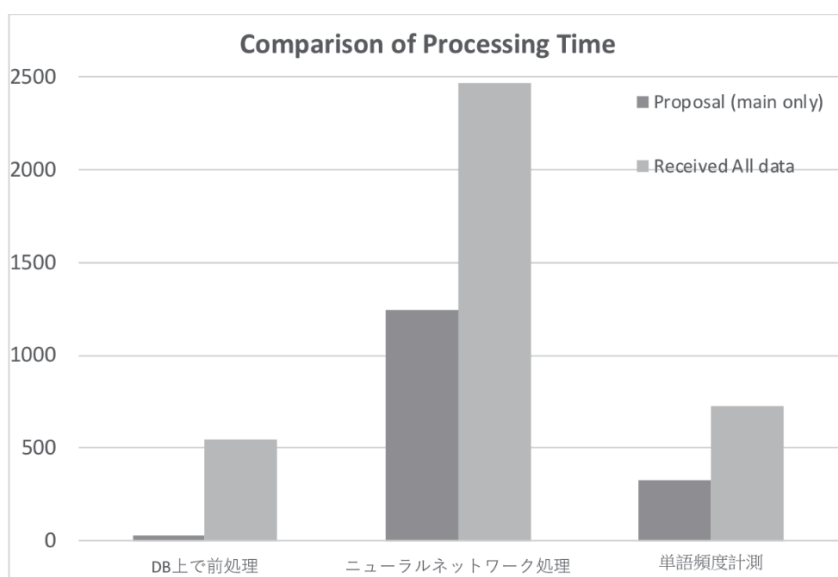


図 5 処理時間の比較

ユーザの取得したメインコンテンツだけでなく全コンテンツ情報を学習した手法(“Received all data”)では、提案手法(“Proposal (only main)”)よりも一部の区間では性能を上回っている。これは、メインコンテンツの抽出が完全でなく一部情報を失った可能性と、メインコンテンツに類似した web 広告などが埋め込まれたことに起因してメインコンテンツの特徴を強調された可能性が検討できる。全てのデータを対象とする場合は、メインコンテンツに加えて、動的に生成されるヘッダ、サイドの広告、SEO(Search Engine Optimization)用のタグが付加される。これら情報は、メインコンテンツに類似するもの、cookie などユーザの web 履歴に依存して動的に付加されている。従って、メインコンテンツのみを対象とした評価はどの区間においても性能が安定しているのに対し

て、全情報を対象とした評価は性能が区間により安定していないと読み取ることができる。

次に、取得した情報からメインコンテンツの抽出の有無による処理時間の関係を図 5 に示す。図 5 では、図 4 を求める過程において、発生した処理時間を比較している。具体的には、DB 上において各サイト・コンテンツから単語抽出の前処理に要した時間、3.1 節で述べたニューラルネットワーク上での文脈学習に要した時間、従来手法の出現単語の頻度計測処理に要した時間を比較している。この差は、単純に処理対象の情報量の差から発生している。そのため、対象となる情報量に比例して影響が大きくなる。

5. 考察

評価結果は、提案手法を利用することで、類似するユーザの興味情報を高い精度で取得できることを示していた。特に、評価ではメインコンテンツの抽出により、安定した性能を有していた。また、比較した語の頻度に基づいた統計的性質を利用する手法では、提案手法と比較して十分な性能を有していない。これは、取得した情報が特定分野、トピックへの偏り(情報の網羅性)が考えられ、出現した単語の頻度という統計値では特徴を表現することができていないと考察できる。つまり、この結果は、情報に含まれる文脈、具体的には単語の語順を学習する提案手法の方が、単語の頻度を用いた手法よりも情報の特徴を表現できているといえる。特に、本稿の評価では、評価データと類似する web 行動履歴のモデルを検索する単純な評価であること、対象となるデータも表 1 に示した 4,500 件程度と多くはないため、統計データとしては不十分であることに起因して性能差が表れたと考察している。

本稿では、2.5 節で示したように、ユーザの取得情報を得るために Web コンテンツからメインコンテンツの抽出を行っていた。この処理は、図 4 および、図 5 で示すように安定した性能、処理負荷が小さくなることから、データクレンジングに相当すると考えることができる。しかし、この処理には、ユーザごとに直近に取得した Web 情報の全取得、解析という工程が新たに必要となる。実システムへ本手法を適用することを検討すると、処理負荷、データ移動のネットワーク負荷、そしてネットワーク内のどのノードにおいて実行するかという新たなプロトコルの設計など解決すべき問題が発生する。また、増大する処理負荷を定量的に検討するために、ランダムに web サイト 3000 件を取得し、2.5 節に示した手法でメインコンテンツの抽出を行なった。このとき、Web ページからメインコンテンツの抽出、情報を示す単語の抽出処理に要する時間とデータ量の関係は、メインコンテンツのデータ量に比例し、処理時間は傾き 0.13 で線形に増加していく。具体的には、80k[Byte]のメインコンテンツで処理に要する時間は 1.1[sec]、140k[Byte]のメインコンテンツで 1.8[sec]程度となる。つまり、本稿の提案手法をシステムで実現するためには、この処理により増加した遅延の改善と、どのノードにおいてユーザ取得情報の抽出をどこで行うのかという検討が必要となる。特定の処理ノード、具体的には各ユーザの web プロキシ的に配置するノードにおいて行うと仮定すると、このノードへ対象のユーザ全てのメインコンテンツ抽出、単語の抽出処理が集中することに起因して負荷が大きくなる。そのため、負荷分散の手法の検討が必要になる。また、各ユーザのノードにおいて、コンテンツ取得時にこれら処理を行う手法も検討できる。しかし、負荷は分散することは可能となるが、現在よりもコンテンツ取得に時間を要することが想定され、ユーザの Web 利用ユーザビリティを損なう可能性がある。

以上のことから、実装の観点から本稿の提案手法を検討すると、どのようなシステムモデルでユーザの web 行動履歴の処理を行うかが提案手法を実現する上での重要な要因となることが予想される。

6. 今後の課題

本章では、考察に基づき、今後の検討すべき課題を明らかにする。

今後、ユーザの興味情報がもっと明らかである Web 行動履歴を学習したモデル構築を検討する。現在は Web 広告をクリックしたユーザは表示された広告に興味があるという仮定を前提としていた。しかし、この前提を進めて Web 広告をクリックしてこの広告の商品を購入するなどのコンバージョンまで到達した、動画映像を最後まで見たなど結果が明らかなユーザ興味情報を学習データとして利用することにより、直近のユーザの興味との関連性、提案手法の有効性の評価を行いたいと考える。

各ユーザの興味の継続する時間を参考文献 [5] に基づいて一律で最長 30 分と仮定していた。これは、状況、人に依存して変化することは容易に想像できる。そのため、ユーザが取得するデータ相関性を調べることで、ユーザ興味の変位、思考の変遷を類推できると考察している。そのため、web 行動履歴間のデータ相関性に閾値を設定することにより、ユーザ興味の変移をトリガとする手法の検討を行う。これにより、学習用データから対

象外データの除去などを期待できるため、提案手法の品質向上を期待することができる。

評価では、対象とするデータはある1日にweb サービスから取得されたweb 行動履歴のみを対象としていたが、今後期間を延ばして評価を検討している。実システムとして適用を想定すると、提案手法はweb 行動履歴となるデータ取得、解析に負荷を要することから、今後これらを想定した提案手法の分散処理手法を検討する。

7. まとめ

本稿では、ユーザが取得する Web 情報に各ユーザの興味情報が含まれていると仮定し、ユーザの web 行動履歴に基づいた興味情報の予測手法を提案していた。提案手法ではユーザの直近 30 分以内に取得した情報の特徴から各ユーザの興味情報を予測する。提案手法の利用により、短期的な興味情報に限定されるが、従来の過去の履歴や、cookie を利用する現在の手法よりも、ユーザの直近の興味に適応可能であること、評価から従来の単語の出現頻度に基づいた統計的手法より高い精度の予測が可能であることを示していた。今後の課題として、提案手法の性能向上のために、興味情報の明らかな学習データを利用したモデル構築、本稿では固定であった興味の継続時間を動的に検出する手法の検討が必要となることを明らかにしている。

謝辞

本研究の一部は、科学研究費基盤研究(C) JP17K01149 の助成にて行われました

[参考文献]

- [1] 株式会社 電通, “2018 年の広告費”, <http://www.dentsu.co.jp/news/release/pdf-cms/2019023-0228.pdf>
- [2] 山口由莉子, Panote Siriaraya, 森下民平, 稲垣陽一, 中本レン, 張建偉, 青井順一, 中島伸介, Web 広告推薦のための長期的・短期的興味を考慮したユーザの潜在的興味分析方式 第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2018), C5-1 2018 年
- [3] 澤田, Siriaraya, 森下, 稲垣, 中本, 張, 中島: Web 広告推薦のためのユーザの興味分析に基づく Web 閲覧予測手法の提案, 第 11 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2019), I2-5, 2019 年
- [4] 久松俊道, 外川隆, 朝日真弓, 生田目崇, EC サイトにおける購買予兆発見モデルの提案, オペレーションズ・リサーチ: 経営の科学, 2013 年
- [5] Watanabe, Y., Ikegaya, Y. Effect of intermittent learning on task performance: a pilot study. J. Neuronet., 38:1-5, 2017
- [6] 広瀬信輔, “アドテクノロジーの教科書 デジタルマーケティング実践指南”, 翔泳社, 2016
- [7] PhantomJS, “<http://phantomjs.org/>”
- [8] Selenium, <https://www.seleniumhq.org/>
- [9] Orkut, B., Hector, G.M. and Andreas P.: Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices, Proc. 10th International Conference on World Wide Web, pp.652-662, ACM (2001).
- [10] 中村健二, 田中成典, 山本雄平, 安彦智史: 共起関係の抽出範囲を考慮した有害情報フィルタリング手法, 情報処理学会論文誌, Vol.54, No.2, pp.571-584, 情報処理学会 (2013)
- [11] 山本 雄平, 中村 健二, 田中 成典, 安彦 智史, “個別の詳細記事抽出のための Web ページ分割手法の提案”, 情報処理学会論文誌 Vol.55 No.2, pp.874-891, Feb. 2014
- [12] MeCab, <http://taku910.github.io/mecab/>
- [13] Quoc Le, Tomas Mikolov, “Distributed Representations of Sentences and Documents”, Proc. of Int. Conf. on Machine Learning, Beijing, China, 2014 vol. 32.
- [14] 徳永健伸, “情報検索と言語処理”, 東京大学出版会, 1999
- [15] D. Blei, A. Ng, and M. Jordan, “Latent Dirichlet Allocation”, in Journal of Machine Learning Research, 2003, pp. 1107-1135.
- [16] topic modelling for humans, “<https://radimrehurek.com/gensim/>”
- [17] TensorFlow, “<https://www.tensorflow.org/>”

(2019 年 7 月 26 日受理)