DEIM2020 P2-16(day2 p89)

Web IndeX システムにおける意外性向上のための関連ページ推薦手法の提案

陳　星瑜† 遠山　元道††

† 慶應義塾大学大学院理工学研究科　〒 223-8522 神奈川県横浜市港北区日吉 3-14-1
†† 慶應義塾大学理工学部情報工学科　〒 223-8522 神奈川県横浜市港北区日吉 3-14-1
E-mail: †chen@db.ics.keio.ac.jp, ††toyama@ics.keio.ac.jp

あらまし　既存の推薦手法は、推薦結果で予想外の結果を出す推奨方法については、あまり研究されていない。意外性推薦とは、ユーザーが思いかけていないが、興味があるかもしれないものを推薦する。本研究ではキーワードが所属しているカテゴリのインターセクトを用いて、意外性を表現できるキーワードとそのキーワードに関するページを推薦する。これによって、無関係や意外性がないものを推薦されてしまうという問題を解決するのは本研究の一つの課題である。Web IndeX(WIX) システムとは、閲覧中の Web ページ内の文章に出現するキーワードを、それに対応する URL へのハイパーリンクへと変換するというシステムである。Web ページの構造化データと文脈情報を使って、推薦されたキーワードのリンク先に対しての必要なページを効率的に取得することができる。本研究では、ユーザーの潜在的な嗜好に基づいて、意外性があるキーワードとリンク先のページを推薦することを目的とした関連ページ推薦手法を提案する。

キーワード　情報推薦，Web IndeX，意外性

1 はじめに

既存の推薦システムは，主に協調フィルタリング、内容ベースフィルタリングとハイブリッドアプローチの三つの手法で実現されている [1]。ただし、実際の世界では、ユーザーは同様の推奨項目に満足するだけではない。たとえば、寿司から醤油を推薦する。この結果は当たり前のことである。ユーザーは、Web を閲覧するときに予期しない推奨事項に興味がある。たとえば、醤油からパンを推薦する。この推奨事項の結果については、ユーザーは醤油で作られたパンが見たことがないので、興味を持つようになる。ただし、推奨結果で予想外の結果を出す方法については、推奨研究でほとんど議論されていない。

本研究では Web IndeX(WIX) における意外性向上のための関連ページ推薦システムを提案する。WIX システムでは、閲覧中の Web ページ内の文章に出現するキーワードを、それに対応する URL へのハイパーリンクへと変換するという既存のシステムである。ログデータから分析されたカテゴリと次クリックするキーワードに合わせて、ユーザーが意外性を思うキーワードを推薦する。従来の推薦システムは各サイト内で行うものが多い、WIX システムでは WIX ファイルに書かれた、全てのサイトが対象になる。本推薦手法は WIX システムにおけるので、推薦できる範囲も全てのサイトになる。この研究では、WIX システムで行う推薦なので、推薦の連続性を表現できる。

意外性を評価する指標は，主観的な要素が多いため，ユーザーがアイテムに対して与えた評価値を用いて定量的に評価することが難しい。今回ユーザーが推薦されたキーワードに対してのクリック率を利用して、ユーザーが意外性があるキーワードに持つ興味を定量化する。他の既存の推薦手法と比べて、処理時間や応答時間も評価する。

本論文の構成は以下の通りである。まず、2 節で Web IndeX システムについて説明する。次に、3 節で関連研究について述べる。4 節で提案する推薦手法のアーキテクチャと処理方式について述べる。5 節で評価について説明した後、最後に 6 節で今後の課題と結論を述べる。

2 Web IndeX システム

2.1 システムの概要

Web IndeX システムとは、キーワードと URL の組み合わせであるエントリの集合を、XML 形式で記述した WIX ファイルを用いて、Web ページ内の文章に出現するキーワードを対応する URL へのハイパーリンクに変換するというシステムである。

2.2 WIX ファイル

WIX ファイルの中で、キーワードは keyword 要素として、それに対応する Web ページの URL を target 要素として記述する。keyword と target を合わせて、エントリと呼ぶ。また、header 要素にそのファイルの概要やメタデータ、作成者のコメント等を記述することも可能である。WIX ファイルの例を図 1 に示す。

2.3 アーキテクチャ

WIX ファイル全てをファイル単位で管理しているシステムのことを WIX ライブラリと呼ぶ。WIX ファイル作成者は、WIX ライブラリを通じて、WIX ファイルをアップロードすることができる。アップロードされた WIX ファイルはエントリ

```
<?xml version='1.0' encoding='utf-8'?>
<!DOCTYPE WIX SYSTEM
"http://wixdemo.db.ics.keio.ac.jp/wixfile.dtd">
<WIX>
  <header>
    <comment>example</comment>
    <description>2020</description>
    <language>ja</language>
  </header>
  <body>
    <entry>
      <keyword>埼玉県</keyword>
      <target>https://www.tripadvisor.jp/
      Tourism-g298175-Saitama_Prefecture_
      Kanto-Vacations.html
      </target>
    </entry>
    <entry>
      <keyword>大阪府</keyword>
      <target>https://www.tripadvisor.jp/
      Attractions-g298199-Activities-Osaka
      _Prefecture_Kinki.html
      </target>
    </entry>
    <entry>
      <keyword>北海道</keyword>
      <target>https://www.tripadvisor.jp/
      Tourism-g298143-Hokkaido-Vacations.html
      </target>
    </entry>
  </body>
</WIX>
```

図 1　WIX ファイルの例

単位に分割され WIX DB に格納される。WIX DB では、WIX ファイルをエントリ単位に分割し、それぞれを RDB のタプルとして管理するデータベースである。

WIX DB 内のエントリ情報をメモリ上に展開し、高速なアタッチを実現するためのオートマトンのことを Find Index と呼ぶ。WIX システムでは、Aho-Corasick 法に基づくオートマトンを構築し、辞書式マッチングを行う [2]。

WIX システムのクライアントサイドは Chrome の拡張機能によって実装されている。ツールバー内のボタンをクリックすると、そのボタンに対応する WIX ファイル内のキーワードが Web ページ中に存在した場合、WIX ファイルの target 要素として記述された URL へのハイパーリンクに変換される。この結合操作のことをアタッチと呼ぶ。アタッチ前後の Web ページを図 2 に示す。

図 2　アタッチ前後の Web ページ

3　関 連 研 究

Web IndeX システムにおいて本研究と同様に推薦を目的としたものとして、「ウェブ資源を利用した Web IndeX における関連コンテンツ推薦システム」[3]、「Web Index におけるキーワード適合型サービス推薦システム」[4] と「ページ適合型 Web Index(WIX) ファイル推薦システム」[5] が挙げられる。ウェブ資源を利用した Web Index における関連コンテンツ推薦システムとは、WIX を利用してウェブブラウジングを行っているユーザーが次に見たいと思われるコンテンツを提示することを目的とした、 WIX システムの追加モジュールである [3]。Web Index におけるキーワード適合型サービス推薦システムとは、サービスの数が増えた時に生じた、ツールバーによる登録個数の制約やサービスを探すコスト増加という懸念事項を解決する推薦システムである [4]。ページ適合型 Web Index(WIX) ファイル推薦システムとは、今後増加が想定される WIX ファイルに対し、閲覧中の Web ページに適したものを容易に取捨選択することを目的とした WIX ファイル推薦システムである [5]。

最初の一つの WIX における推薦システムとしては、関連しているキーワードしか推薦されていない。現在推薦手法がたくさんある、最近はキーワードの間で Relation Type と Relation Value を付けて、キーワード間の関係を決めて推薦するという推薦手法が研究された [6]。残りの二つの WIX における推薦システムとしては、ツールバーで探すコストを減るためという目的した推薦システムである。推薦リストは通常、一番上のアイテムが最も正確であるという印象を与える。ただし、1 つ以上のアイテムが他の推奨アイテムと著しく異なる場合、ユーザーの興味と好奇心を刺激する可能性がある。また、推奨の品質を向上させることができる [7]。この研究では、キーワード間の関係だけではなくて、キーワードから、カテゴリを使って、意外性があるキーワードと対する Web ページを推薦する。WIX ファイルを利用して、Web ページが繋がれるという WIX システムの特徴を使うのは本研究の推薦手法を WIX システムにおける目的である。

現在、ウェブサイトを見る時、横に推薦されるニュースや記

事サイトが多いが、興味がないものも多い。コンテンツの人気とユーザの関心の両方が時間とともに頻繁に進化するため [8]、従来の推薦手法がよく推薦できない。本研究では、カテゴリと次にクリックするキーワード情報を使って、推薦キーワードを生成して、関連ページを推薦する。ログデータはユーザの好みをよく反映するので、カテゴリを決めるデータをログデータをする。今回の研究では、ユーザーの WIX システムの使用履歴を使って、Web ページを推薦する。それぞれのユーザーの WIX システムログデータを使うことで、推薦のパーソナライズ性も増加できる。WIX ファイルは図 1 のように記入しているので、キーワードをクリックするうち、キーワードと URL をログというテーブルをサーバ上で保存することができる。自分のログデータによって、推薦するならば、他のユーザに関わらないので、従来の推薦手法の欠点を改善できる可能性がある。

4　意外性向上のための関連ページ推薦

4.1　概 要

本論文で提案する推薦手法のイメージを図 3 に示す。従来の WIX システム機能を保持する上、本研究の推薦手法を加えた。Web ページ上のキーワードにマウスオーバーして、横に意外性があるキーワードを推薦するポップアップメニューが見える。直接本文中のキーワードをクリックすれば、本来の WIX システムとして使える。本論文で提案する推薦システムの概要を図 4 に示す。本研究では、本来の WIX システムの機能を保ちながら、本推薦手法を提案した。ポップアップメニューで現れるキーワードも WIX システムと同様にハイパーリンクになっているので、クリックしたら、そのキーワードに対応したページにアクセスできる。これは関連ページ推薦である。直接のページへの推薦ではなく、キーワードを使うことにより、ユーザーが複数の候補から選ぶという自由度も本研究一つの目的である。

図 3　WIX における推薦手法

4.2　推 薦 条 件

本研究では、今見ているページ中のキーワードとログデータから分析されたカテゴリに合わせて、次のキーワードを推薦す

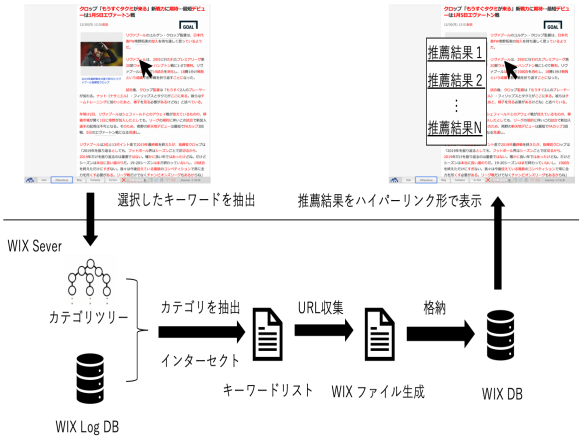

図 4　推薦手法のアーキテクチャ

る。推薦条件の要素は、簡単に言うと、キーワードとカテゴリ二つである。

もし、単なるキーワード間の距離や類似度で、新しいキーワードを推薦するなら、同じキーワードが複数のカテゴリに所属する場合が多い。日本版の Wikipedia の例としては、「リヴァプール FC」というキーワードに対象として、所属しているカテゴリは「イングランドのサッカークラブ」、「リヴァプールのスポーツ」、「リヴァプールの組織」、「プレミアリーグのサッカークラブ」、「リヴァプール FC」、「1892 年設立のスポーツチーム」合わせて六つのカテゴリがある。この六つのカテゴリ全て単なるのカテゴリではなく、上位や下位カテゴリも含まっている。「イングランドのサッカークラブ」は「リヴァプール FC」の上位カテゴリになっている。

本研究では、カテゴリを使って、推薦されるキーワードの範囲を小さくする方法を考えている。sports というカテゴリ一つで、サラー選手を確定することはできない。この問題に対しては、Wikipedia のようにカテゴリツリーを構築する。大きなカテゴリ下で、小さなカテゴリを作って、階層的にカテゴリを分解する。これによって、ユーザーが興味に対して、推薦結果をより強く制約する。一つのキーワードに対して、複数カテゴリに所属することが可能なので、それらのキーワードがカテゴリ間の関連キーワードになる。両カテゴリ間の意外な関連をあらわす項目として興味深い項目である [9]。キーワード k の所属するカテゴリ集合を C_{key} を定義し、集合 C_{key} 中の任意のカテゴリを C_k を定義する。ログデータから抽出するカテゴリの集合を C_{log} を定義し、集合 C_{log} 中の任意のカテゴリを C_l を定義する。意外性を表現できるキーワードの集合 E を式 1 で定義する。

$$E = C_k \cap C_l \{C_k \subseteq C_{key}, C_l \subseteq C_{log}\} \quad (1)$$

E が含まれている要素が多いほど、カテゴリ間の関係が近い、所属関係のカテゴリの可能性も大きい。集合 C_{key} が含まれている要素が多いほど、知らないカテゴリがある可能性が大きいので、そのカテゴリから推薦されたキーワードも意外性の表現率が高い。Wikipedia の例として、日本の有名なサッカー選手本田圭佑が所属している一つのカテゴリは「日本・カンボジア」

関係。これに基ついて、明石康さんを推薦したら、意外性を表現できるかもしれない。本研究では、どのキーワードをどのカテゴリに所属するか Wikipedia のように事前に決める。ユーザーが Web ページ中のキーワードにクリックすると、そのキーワードに所属しているカテゴリとユーザーのログデータから分析された関心があるカテゴリを式 1 のように集合 E を取得することができる。推薦結果 R を式 2 のように一番小さい集合 k 中で生成する。

$$R \in min(E) \setminus k \tag{2}$$

4.3 ログデータに基く URL の分析

ログデータ中の target で保存してある URL を分析して、図 5 のようにキーワードセット [10] ような形にする。

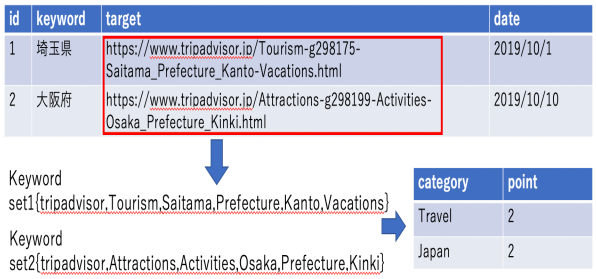

図 5　URL 分析

URL は、最初に「:」、「/」、「?」などの一般的な区切り文字のセットによっていくつかの汎用 URL コンポーネントに分解される。 各コンポーネントは、「&」、「=」、「」、「-」などの別の区切り文字セットによって、いくつかのサブセクションにさらに分割できる [11]。URL は分解に基づいた一連のキーと値のペアで簡単に表すことができる。本研究では、この一連のキーと値のペアをキーワードセットの形でそれぞれ保存して、トレーニングサンプルに基づいて、用意してあるカテゴリツリー中のカテゴリに分類する。トレーニングサンプルについては、どのカテゴリでどのキーワードが含まれるという情報を記入してある。トレーニングセット T を式 3 で定義する。ウェブページの URL をパターンと照合することにより、ユーザーが閲覧したウェブページを分類することができる。一つの URL が複数のカテゴリに所属場合もある。ポイントが高いカテゴリを上位からランキングする、ユーザーが一番興味がある分野を求める。

$$T = \{(k, C_{key}) \,|k \in C_{key}\} \tag{3}$$

ログデータを取得する方法については、処理時間を短縮するため、三つのセッション破棄戦略を考えている。それは頻度が最も低い記録を取らない戦略、入れた順で先に入れた記録を取らない戦略と時間枠を決めて、その時間より前の記録を取らない方法 [12]。頻度が最も低い記録を取らない戦略に対しては、本研究のリアルタイムでの推薦更新という機能に適応していない。新しい入ってくるログデータを取れないので、推薦更新してもユーザーの既存の関心を拡張する能力が向上できない。本研究は入れた順序で記録を取る戦略と時間枠で記録を取る方法を使用する。

4.4 WIX ファイルの自動生成

WIX システムのユーザーがユーザー自身の興味に応じて、キーワードに関連する URL に容易にアクセスできるようにするために、WIX ファイルの自動更新機能を加える [13]。本研究によって提供される WIX ファイルは、推薦されたキーワードから、最新のコンテンツを含む Web ページの URL を取得して、WIX ファイルを生成する。WIX システムでアタッチを行うことにより、ユーザーは推薦されたキーワードに関連する最新の Web サイトに簡単にアクセスすることができる。

推薦されたキーワードに、Web から関するページを取得することがこの研究の一つの課題である。本研究は Web ページの構造化データと文脈情報を使って、キーワードから Web ページを取得する。埋め込まれた構造化データがある Web ページに対して、構造化データを使って、それぞれを推薦されたキーワードに対応することができる。構造化データがないページに対して、自然言語処理を使って、タイトル中に推薦されたキーワードが含まれているページを推薦する。

4.5 リアルタイムでの推薦更新

WIX というシステムは WIX ファイル内で記入しているキーワード全部ウェブ上でハイパーリンク形式で表現するので、ユーザーが興味を変えたとしても、自分の新しい興味のキーワードをクリックするうち、推薦される結果もログデータに応じて、新しいキーワードに変わる。リアルタイムでのページ推薦を図 6 に示す。リアルタイムでの推薦更新により、ユーザーの既存の関心を拡張する能力も向上できる。

図 6　リアルタイムでのページ推薦

4.6 意外性の表現

本研究は意外性向上を目的とした推薦手法なので、意外性を表現する方法を説明する。本研究の意外性は、無関係なものや関連が近いもの、どちらでもない。無関係なものと関連が近いものの間にあるものである。意外性がありすぎると、元のキーワードから離れたら、意味がなくなり、無関係になる。意外性や無関係なものの定義は人によって違うことも多い。例えば、ユーザーが料理のサイトを見る時、イタリアというキーワードをクリックして、推薦されたページはイタリアサッカーチームのページの場合、サッカーに興味がないユーザーとしては、これは無関係なものだと思う。もし、このユーザーはサッカーファンなら、この推薦結果としては、無関係なものより、意外

性がある推薦だと思うかもしれない。このような観察から、本研究の推薦条件中に、ログデータから分析されたカテゴリを使うことにした。無関係なものを式 1 のように定義すると、E = 0 の時は推薦結果が無関係なものになって、C_k=C_l の場合だと、関連が近いものが推薦される。意外性推薦とは、ユーザーが思いかけていないが、興味があるかもしれないものを推薦する。本研究はカテゴリツリーを構築し、キーワードが所属するカテゴリ間をインターセクトして、推薦するキーワード得る。

5 評　　価

本研究は、意外性向上を目的とした推薦手法のために、意外性について主に評価する。意外性を評価する指標は，主観的な要素が多いため，ユーザーがアイテムに対して与えた評価値を用いて定量的に評価することが難しい。今回はユーザーが推薦されたキーワードに対してのクリック率を利用して、意外性があるキーワードの有用性を定量化する。本研究では、ユーザーによるキーワードに対するワンクリックでたくさんの処理が含まれるため、応答時間も評価の一つになる。

5.1 有　用　性

推薦された結果に対し、一つだけを意外性があるキーワードにして、他の推薦結果に対しては、普通に関連度が高いキーワードを推薦にする。意外性があるキーワードにクリックした回数を S とし、推薦されたキーワードにクリックした回数を A とする。

$$\text{クリック率 P} = \frac{S}{A} \tag{4}$$

$$\text{有用性} = \frac{1}{n}\sum_{i=1}^{n} p_i \tag{5}$$

クリック率を式 4 のように定める。ユーザーが意外性があるキーワードにクリックした回数を統計して、そのキーワードがユーザーに対しての有用性を式 5 で定義する。

5.2 応答時間の比較

本研究は WIX システムにおける推薦システムなので、本来の WIX システムと比べる必要がある。本来の WIX システムに比べ、応答時間の差が出ない場合、本推薦手法の処理時間は十分に短いと言える。本推薦手法で選んだ推薦条件はカテゴリとキーワードなので、最近研究されだキーワード間の関係を使った推薦手法との比較も必要になる。応答時間の比較の結果によって、本研究で使った推薦条件を客観的評価することができる。

6 今後の課題と結論

本論文では、いつも変わるユーザーの興味に対し、閲覧中の Web ページから意外性があるページを推薦する手法を提案した。WIX 上で本提案手法を加えると、一つのキーワードに対し一つのリンク先という単調さを改善できるので、ユーザーの自由度や得る情報量が増加することが期待できる。推薦システムを WIX のログデータを利用するため、他のユーザーに関わらないので、パーソナライズ性も向上することも期待できる。WIX と合わせて、推薦の連続性も向上することが期待できる。推薦されるページが意外性を持つため、ユーザーが既存の関心を拡張する能力も向上することが期待できる。提案手法としては十分なものとは言えないため、さらなる他の意外性に目的した推薦手法による比較と評価が今後の課題である。

文　　献

[1] Santosh Kumar, Varsha," Survey on Personalized Web Recommender System", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.10, No.4, pp. 33-40, 2018.

[2] 石崎文規, 遠山元道. 大規模 Aho-Corasick オートマトンにおける追加更新手法の提案. データ工学ワークショップ，DEIM2012. 2012.

[3] 越島亮介, 遠山元道. ウェブ資源を利用した Web Index における関連コンテンツ推薦システム. DEIM2015. 2015.

[4] 井上明莉咲, 遠山元道. Web Index におけるキーワード適合型サービス推薦システム. DEIM2018. 2018.

[5] 新里匠, 遠山元道. ページ適合型 Web Index(WIX) ファイル推薦システム. DEIM2019. 2019.

[6] Xin Xin, Xiangnan He, Yongfeng Zhang, Yongdong Zhang, Joemon Jose, Relational Collaborative Filtering: Modeling Multiple Item Relations for Recommendation, Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, July 21-25, 2019, Paris, France

[7] Ge, Mouzhi, Carla Delgado-Battenfeld and Dietmar Jannach. “Beyond accuracy: evaluating recommender systems by coverage and serendipity.” RecSys ’10 (2010).

[8] Wu, Qingyun, Huazheng Wang, Yanen Li and Hongning Wang. “Dynamic Ensemble of Contextual Bandits to Satisfy Users’ Changing Interests.” WWW (2019).

[9] 野田陽平, 清田陽司, 中川裕志. 意外性のある知識発見のための wikipedia カテゴリ間の関係分析. 第 20 回セマンティックウェブとオントロジー研究会, 2009.

[10] Bharti, Pooja M. and Tushar J. Raval. “Improving Web Page Access Prediction using Web Usage Mining and Web Content Mining.” 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)(2019): 1268-1273.

[11] Yang, Yiming, Lei Zhang, Guiquan Liu and Enhong Chen. “UPCA: An efficient URL-Pattern based algorithm for accurate web page classification.” 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)(2015): 1475-1480.

[12] Bharti, Pooja M. and Tushar J. Raval. “Improving Web Page Access Prediction using Web Usage Mining and Web Content Mining.” 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)(2019): 1268-1273.

[13] 大島拓也, 遠山元道.Web Index における配信型コンテンツを利用した自動ライブラリ更新システムの提案. DEIM2016. 2016.

[14] 図 2、図 3、図 4 で用いたサイト『クロップ「もうすぐタクミが来る」新戦力に期待 最短デビューは 1 月 5 日エヴァートン戦』閲覧日時 2019 年 12 月 30 日 https://headlines.yahoo.co.jp/hl?a=20191230-00010010-goal-socc

[15] 図 6 で用いたサイト『大物買い一転、今夏は 10 代 2 人。リバプール「移籍委員会」のネクストステージ』閲覧日時 2019 年 10 月 28 日 https://headlines.yahoo.co.jp/hl?a=20191230-00010010-goal-socc