論文内容の要旨

A Study on Intelligent Web Site
-Towards a New Generation of Adaptive Web Portals-
（インテリジェントウェブサイトに関する研究
-新世代適応型ウェブポータルの構築に向けて-）

ベラスケス　シルバ　ホアン　ドミンゴ

第１章は、人々の通信手段を改革する社会現象としてのウェブ、ワールドワイドウェブを紹介している。特に、ウェブサイトで実験されている構造変化や、使用者の検索行動の変化によるコンテンツ変化についての調査に焦点をあてる。今日ウェブサイトは、使用者の直接接触を許す静的情報を持つサイトから連続的・定常的情報を持つサイトに発展している。

この発展の中で、ウェブサイトの設計者は、サイトの構造とコンテンツについて、使用者が探す情報を見付けやすくするような設計を常に試みている。しかしながら、現実はそれほど簡単ではない。多くの場合、所望の情報がウェブサイトに収蔵されているにもかかわらず、サイト構造がこれを隠してしまっている。事実、いくつかのサイトでは同一のページの中に収蔵されるコンテンツが多すぎる。構造の悪さと相まって、このことが使用者をして所望の特定コンテンツが見つけだせない「ハイパースペースでの迷子感」を感じさせることになる。このことは使用者によって異なり、ある時ある使用者には全く訳の分からないウェブサイトになっていても、別の使用者には非常に明解なウェブサイトになっている場合がある。しかしながら、ウェブサイトは構造上大多数の使用者に明確なものでなければならない。

本論文では、ポータルを作ることにおける次代の改革を支援するために、ウェブサイトの構造とコンテンツをいかに改革するかの研究に焦点をあてる。このようなウェブサイトを知的ウェブサイト、つまりこのウェブサイト使用者の検索行動の解析結果によりウェブサイト構造とコンテンツを改善可能なウェブサイト、と呼ぶ。ここで使用者とは、その使用者の個人情報が与えられていない、頻繁ではないウェブサイトの利用者を意味している。

第２章では、主要なデータ源であるウェブログファイル内の情報をもとに、使用者のウェブサイト検索行動を解析する。ウェブサイトのトラフィックにもよるが、使用者のクリック毎にウェブログレジスターはページ内に記録を残すため、不必要な情報も含め数百万にのぼる記録を持っており、ウェブログファイルの解析を非常に複雑化している。

特定された利用者や顧客（これらの人の検索活動においては性別・年齢・前回の検索記録などの付加的な情報が得られており、また検索活動の全貌を把握する方策をもっており、つまり各利用者の検索活動を特定することができるので）に比較して、不特定使用者の検索行動解析は、より複雑である。ウェブログデータからは、使用者の検索行動の推測のみが可能である。

使用者の検索行動は、コンテンツと検索連鎖と各ページでの滞在時間によって性格付けられると想定される。これらの情報は、ウェブログファイルおよびウェブサイト自体、つまりウェブサイトのテキスト記録から抽出可能である。

第３章は、ウェブデータ（ウェブログ、ウェブ頁、ウェブリンク）より知識を抽出するために使用されるデータベースにおける知識発見手法（ＫＤＤ）を導入する。まず初めに、使用者の検索過程を再生しウェブサイトから必要な情報を得るためのクリーニングおよび前処理アルゴリズムを開発する。これらのアルゴリズムは、相関データベースエンジンの能力、つまり表やインデックスや素材別ビューといった加速オブジェクト、を使用して創られる。これらのアルゴリズムで得られた結果はウェブデータから抽出された情報の蓄積用に準備された情報蓄積器に収蔵される。蓄積器は、ファクトテーブルが要約情報を収蔵し、次元テーブルがクウェリーを実行するに必要なメタデータを収蔵する、スターモデルを適用するデータマート技法を用いて実装される。

第４章は、ウェブデータの前処理およびクリーニングに使用する技術を導入する。ウェブコンテクストマイニング（ＷＣＭ）とウェブ使用法マイニング（WUM）の背後にある基本概念を組み合わせて使用者検索行動を解析する新しい方法を提案する。この方法はＷＣＭ，WUM両手法を補足するビジョン

を与えている。WUMでは、使用者のブラウジング行動を理解できるが、使用者がどのコンテンツに興味を持ったかを知ることが出来ない。後者の解析はWCMを用いれば可能となる。

結合情報「使用者行動ベクトル」を定義する。これは、見に行ったページの連鎖とその検索行動の中でそのページを見ていた時間の百分率とを対にして構成される。使用者検索行動に対する新しい類似性尺度が導入され、類似検索行動を群別するために開発されたアルゴリズムで使用される。これらのアルゴリズムの内の１つは、使用者行動のような時間的連鎖に対応するデータを処理するときに有用な性質である、投射空間の連続性を維持可能な利点を持つトロイダル位相を持つ自己組織化地図アルゴリズムと同等の働きを持つ。

「ウェブサイト鍵語」という概念が導入され、使用者にとってページをより魅力的にする語あるいは一群の語、つまり使用者にとって最も面白いテキストコンテンツ、を定義する。これらの鍵語はウェブサイトにおける使用者の基本的用語嗜好を与えるものである。

一群のウェブサイト鍵語を設定するためには、使用者の検索行動の中で最も重要なページの選定をする必要がある。ページの重要性は使用者がそのページを見ている時間に相関するという仮定のもとに「重要ページベクトル」を導入する。このベクトルは使用者行動ベクトルを時間要素で並べそのサブセットを検索行動における最重要ページに選定する。

テキスト嗜好が同じ使用者を発見するためにクラスタアルゴリズムも使用する。次に重要ページベクトルの比較のために新しい類似尺度を定義し、得られたクラスタからウェブサイト鍵語を抽出する手法を開発した。

クラスタ識別は主観に依存するため、つまりクラスタから何を理解するかによるので、時々困難な過程となる。しかしながら、どのような時に一群のポイントがクラスタと考えられるかを推定する２つの手法を与えた。核関数では、隣り合わせになったデータポイントの影響が考慮される。密度関数では全データポイントの影響の和が必要である。両方の場合ともクラスタ識別は現在検証中の特定の問題に依存する値を持つパラメータを使用する。また、クラスタ理解は主観的な問題であることから、クラスタの認証と否認を対象とな ら応用分野（例えばビジネス）の専門家に依頼し、新しい解釈を得ることが有益である。

類似の使用者の行動は同じクラスタにグループ化されるという仮定のもとに、認証されたクラスタは未来状況の予測に使用される。この情報をもとに、各新着使用者を特定のクラスタに割り当て、その使用者の行動を予測する。

第５章では、ウェブデータから抽出された知識を獲得し、維持するフレームワークを導入する。ウェブデータから抽出された情報と、クラスタ化技術を適用した後に発見された知識とを蓄積するために、２種の構造が開発された。

最初の構造は情報蓄積器であり、そのコンテンツはクラスタ化アルゴリズム源として使用される。発見された知識は蓄積のためにより複雑な蓄積器を必要とする。その実現のために、パターン蓄積器とルール蓄積器で構成される知識ベース（ＫＢ）が設計された。

識別されたクラスタから、使用者ブラウジング行動の重要なパターンと使用者のテキスト嗜好を抽出することが可能である。ビジネス専門家の支持のもと、以下にパターンを使用するかに関する一群のルールが「if-then-else」プログラム命令として創られる。パターンとルールはパターン蓄積器とルール蓄積器にそれぞれ収蔵される。両蓄積器が使用者行動とウェブサイトにおけるその使用者の基本的嗜好に関わる知識ベースを構成する。

ここで導入されたフレームワークは２種類の利用者を持つ。２種類の利用者とは、彼等自身の好みに基づいて個性化された実時間推奨案内を受ける個人利用者とウェブサイトの構造とコンテンツの変更に関わる非実時間推奨を受けるウェブマスターである。

第６章は、本論文で開発された理論、モデルおよびアルゴリズムの実在ウェブサイトからのデータ、つまりウェブログとウェブページへの応用を示した。

幾人かの研究者が、複雑なウェブサイトからの実データへの応用を検証することが難しいことを指摘している。伝統的に、これらのサイトは商業会社に所属しており、方針上これらのデータの開示には否定的である。その結果通常取られる方法は、公共的ウェブサイト、例えば大学のコースウェブサイトを活用することである。

本論文では、２つの実ウェブサイトを活用できた。第一のサイトはチリ大学の産業工学科における連続教育プログラムウェブである。第二のサイトはチリで最初の仮想銀行に所属するウェブで、ここでは物理的支店窓口がなく、すべてのトランザクションが最初から電子的に作られている。解析期間（2003年１月から３月まで）の間におおよそ 800 万の生データが登録された。

教育ウェブサイトでは、2002 年 8 月から 11 月までのデータが獲得された。2002 年末にウェブサイトに非実時間変更が行われた。2003 年の同じ期間にもデータを獲得し、両期間とも約 6000 の比較対照となる検索行動が観測された。2003 年のデータは、変更により、３ないし４ページの検索活動が 2002 年に比較して 11%増加し、１ないし２ページの検索行動を 27.4%減少し、また検索行動の時間を長くし、最も見る回数の多い 20 ページを見る使用者を増加させた。このことは、このウェブサイトの構造とコンテンツが使用者の興味を増加させていることを示している。

銀行ウェブからは、2003 年 1 月から 3 月までの 800 万に及ぶ生データを処理した。抽出された情報は開発されたウェブマイニングアルゴリズムへの入力として使用され、このアルゴリズムの出力は使用者ブラウジング行動とその使用者の文字嗜好に関する重要パターンとなる。ウェブサイトがこの銀行の中枢であることから、ウェブサイトに対するすべての変更は銀行首脳部の承認事項であるために、承認までに時間がかかり、現在変更中である。しかしながら、活用可能なウェブログから実時間推奨変更の効率性を検証することが出来た。提案システムと対応するドメイン知識に立脚した実時間推奨案内を提供するルールを作り出すために３ヶ月間（2003 年 1 月から 3 月）のログファイルから最初の 70%を選択した。使用者の対応期間は 1 月 1 日から 3 月 10 日である。

ルールはウェブログファイルの残りの 30%のデータに適用され（3 月 11 日から 3 月 30 日まで）、以下の手順で効率性が検証された。実時間推奨案内はウェブサイトの現使用者により指示される同じウェブサイト内の一群のページである。推奨案内を使用者が受容したか拒否したかを知るために、指示されたページのコンテンツと使用者により現実に選択されたページとが比較された。提案された手法を使用すると１ページのみが推奨され、50%を少し越える場合においてこのページが受容されたという結果が示された。これは、扱っているページ数が多く、ページ間の多くのリンク数も多くかつ数回のクリックで去ってしまう確率が高い使用者の多い複雑なウェブでは、非常に成功した推奨であるとビジネス専門家は評価している。

上記に付け加えて、検索行動過程で推奨されたページが現実に含まれている場合も含めると受容率はさらに高くなることを述べておく必要がある。これはログファイルに蓄積された過去の使用者を比較しており、提案の推奨を全く受けていない使用者の検索行動のみしか解析できないことに起因している。

第 7 章は、本論文の主要結論を述べている。ウェブサイトにおける使用者のブラウジング行動とその使用者の嗜好を解析するための開発されたアルゴリズムとモデルから、知的ウェブサイト提案は有効であることが結論づけられる。

本論文の主要成果は下記のとおりである。

- ウェブサイトにおける使用者検索行動理解のための類似尺度の新提案
- トロイダル位相による自己組織化地図に基づくウェブマイニングクラスタアルゴリズムの新提案
- 実時間推奨案内のための知識の獲得保持の新方式
- ウェブサイト鍵語発見のための新手法
- 実時間推奨案内と非実時間推奨ウェブサイト構造とコンテンツの新手法

残された課題としては、ウェブサイト鍵語の概念をページ中の他の要素、例えば静止画像、動画像、音響などに拡張可能である。例えば、使用者に実際にデータベースを見に行くようにし向けるようなコンテンツ、ウェブサイト鍵オブジェクトに関する研究を提案する。使用者とのより良い関係を実現する目的で、提案による変化の効率性やインターフェース推奨における関係する使いやすさ要素を検証するためには、実験を計画することも重要である。

以上