

ユーザ行動モデル構築のためのデータ解析システムの構想

L^AT_EX Style file for manuscripts of JSAI 2003

本村 陽一^{*1} 中村 嘉志^{*2} 西村 拓一^{*3} 幸島 明男^{*3} 和泉 憲明^{*3} 松尾 豊^{*3}
 Yoichi Motomura Yoshiyuki Nakamura Takuichi Nishimura Akio Sashima Noriaki Izumi Yutaka Matsuo

^{*1}産業技術総合研究所 デジタルヒューマン研究センター/JST CREST
 AIST DHRC/JST CREST

^{*2}産業技術総合研究所 サイバーアシスト研究センター
 AIST CARC

^{*3}産業技術総合研究所 サイバーアシスト研究センター/JST CREST
 AIST CARC/JST CREST

We propose a data handling system for user's action analysis. In general, history data of user's behavior is stored with fixed representation in each attributes of data bases. However, sometimes we need more flexible handling and matching scheme even if data and query have different representation in attributes. We implement a middle ware to convert and match different representation. This software is used in this annual conference of JSAI, and it can be applied for data augmentation in statistical learning probabilistic user and task models.

1. はじめに

我々は人間中心の情報アクセス技術の一つの方法として、実環境におけるユーザの行動履歴に基づいて、ユーザを適切に支援する情報提供サービスの実現を目指している。とくに全てのユーザに同じ内容の情報サービスを提供するのではなく、個々のユーザの嗜好性や特徴、行動履歴の違いに応じて、個別化した情報サービスの提供に焦点を当てている。

そこではユーザが環境内で行動し、様々に異なる情報サービスを利用するうちに、適合性フィードバックによってその好みや傾向などをシステムが理解しユーザの意図や要求などを予測できるような仕組みが重要である。

そのためにはユーザの行動履歴データを解析することが必要であり、これまで対話的な情報検索システムにおける検索要求の表現の違いや結果から受ける印象の差などの個人の主観に起因する違いをモデル化するための考察 [1] や、対話的なタスクをユーザとの対話履歴から柔軟に変更できるようなタスクモデルの提案 [2] などを行ってきた。

これらのユーザモデルやタスクモデルに関して統計的な学習を行う、つまり大量の統計データからモデルの確率パラメータを学習することによって、モデルはユーザに適応し、適切なタスクを実行することが可能となる。そこでのポイントは性能の良い学習アルゴリズムと、質の高い大量の統計データである。いかに良いモデルと学習アルゴリズムがあっても、十分な量の質の高いデータがなければ適切な学習結果は得られない。

この時、データをどのように表現するかは非常に重要な問題である。例えば、ある特定の場面について観測されたデータが、実はそのユーザにとっては固有だが、他のタスクなどにも一般的に成立する個性を表していることもある。しかし、データの属性表現が固定的であると、観測時と同じ表現型式で表される、ごく限られた場面にしか適用されないことになる。例えば個々の情報サービスごとにユーザの動作履歴データを収

集するのでは、個々の情報サービスのそれぞれについて十分なデータ量が必要となり、それが実際の統計的学習に基づくアプローチを実現する上での大きな問題であった。観測されたデータが、特定の場面でしか適用されず、データの持つ情報が十分に活用されない、複雑なモデルのためには非常に大量のデータが必要となる。そのため、実際に得られたデータだけでは、十分に学習できないという問題も生じるのである。

そこで本研究ではユーザやタスクに適応する確率モデルの統計的学習の性能を改善するための一つのアプローチとして、学習アルゴリズムの改良ではなく、データの品質と量の問題に注目し、観測時に得られたデータをできる限り一般的に解釈して、適用できる範囲を拡大することを考える。まず表面的には異なる属性表現をとるデータであっても、ある観点から見た時に共通する性質を持っている場合には、同値類としてマッチングがとれるような検索を可能とする XML を活用したデータベースシステムを開発した。またこれとユーザの移動履歴を計測するセンサネットワークを利用したイベント支援システムの概要を紹介し、本全国大会におけるユーザの行動履歴を収集し、データを解析するシステム全体の構想について述べる。

2. イベント支援システムにおける行動履歴の収集

インターネット上の情報サービスと同様に、実空間における情報提供サービスも容易に実現できるようになってきた。例えばユーザはサービスが提供されている場所に行き、そこで自然な要求行動をとることで、そこで適切な情報サービスを受ける。本学会においては、無電源の情報端末である CoBIT[3](ボタン付 ID 発信器搭載型 CoBIT 名札タイプ) を持つ聴講者が、デジタルポスター展示や、セッションの会場に行き CoBIT を環境側に設置した光源に向けて音声が得られる。この時同時に個々の CoBIT に割り当てられている ID が検知され、聴講していた場所、時刻、内容に関する情報が記録される。CoBIT を参加者に配布する時に ID と参加登録者対応をとることで、参加登録時に入力した個人情報と結びつけることができる。さらにインターネット検索から得られる人間関係ネットワーク [4] を利用することで大会での各研究発表と、それに興

連絡先: 本村 陽一, 産業技術総合研究所 デジタルヒューマン研究センター, y.motomura@aist.go.jp, URL:http://staff.aist.go.jp/y.motomura/, 〒135-0064 東京都江東区青海 2-41-6 産総研 臨海副都心センター

味を持った研究者の研究テーマとの関連性なども解析することが可能となる。(これらのイベント支援システムの詳細については [5] を参照してほしい.)

3. 行動履歴の格納と検索時の問題

以上に述べたユーザの行動としての位置情報や時間情報, また情報サービスの内容を記録する時の表現形式と, データを検索・解析する時に用いる表現形式が異なる場合がありえる. また新たなセンサや情報サービスを追加した場合でもユーザに関する情報としてはそれまでのデータとの整合性を取る必要もある. こうしたデータ表現の差異を扱うためには XML の利用が有望であり, 異種のセンサの表現を統一的に扱うために XML を利用した試みもすでにある [6].

ここでは, もっとも簡単な例として, 位置情報, 時間情報についてみてみよう. 記録時には, センサの許す限り詳細な表現を用いて, 位置情報, 時間情報をデータベースに登録する. 例えば, CoBIT 端末であれば, ID 検出器を設置した位置から検出器の有効範囲内のどこかにユーザが居ることがわかり, また無線 LAN による位置検出の場合には, 無線基地局の設置場所の周辺の無線の有効範囲, あるいは信号強度による三角測量に基づき詳細な位置が検出できることもある. 時間については記録時刻をタイムスタンプとして付与しておけばよい.

一方, 記録されたデータを検索し, モデルを構築したり, アプリケーションが利用する場合には記録時の表現形式・粒度と整合しないことがある. 例えば, 各パネルにおける位置情報と正確な時刻を記録した後に, 「パネルを含むそのフロア全体に居た参加者」を検索したり, 「午前の関連セッションに参加していた参加者」を検索するような場合である. 時刻の場合だけであれば, 従来の SQL データベースでは SQL 文の中で時刻の上限, 下限を指定するなどして検索を行うこともできるが, 検索する側が属性としてデータベースへの格納型式を知って, 適切に解釈しなければならない. また複数の粒度の異なるセンサにより検出した位置情報の場合であれば, 記録する時点で異なる粒度でデータが格納されるために, 検索時に表現形式の整合性をとる操作が従来の SQL 文を用いる方法では繁雑になるため拡張性が損なわれる.

4. 行動履歴データ解析のためのミドルウェア

先に述べたデータ表現の整合性の問題を解決するために, 我々は XML 表現によって異なる粒度の位置情報, 時間情報を適切に変換する機能をミドルウェアとして実装し, XML データベースと連携して利用する. これによって, 検索時にはアプリケーションが必要とする粒度で検索要求を生成し, ミドルウェアが異なる表現の整合性をとり, 該当するデータを検索結果として返す. これによりモデルを学習するためにデータの頻度を必要とする場合に, 細かい粒度ではデータ量が不足し, 頻度が得られない場合でも粗い粒度で問い合わせを行うことで, 十分な量の検索結果から頻度を求めることができる. この時, 従来のようにモデルの表現を変える度に元のデータを全て変換する必要はない.

また, 格納されているデータの粒度が粗い場合 (例えば「時刻 t にフロア A に居た」というデータについて「フロア A のうちパネル B に居た参加者」を検索要求とする場合) には該当する可能性のあるデータと, その確信度を合わせて結果として返す. これにより, 完全に合致するデータがない場合でも可能性のある候補を結果としてアプリケーションが利用することができる.

さらにこうした表現の差異をミドルウェアで吸収する仕組みを利用することでユーザの行動の解釈が複数ありえる場合 (例えば, 展示物に接近した, 単に移動中通りすぎた, 知合いに近寄ったなどの複数の解釈がありえる場合) などについても同じ枠組みで取り扱うことが可能であり, アプリケーションやユーザモデル, タスクモデルの構築の際に有効であると期待できる.

5. おわりに

ユーザ行動モデルを構築するためのデータ解析のために行動履歴を収集し, 粒度の異なる検索要求に適切な結果を返すミドルウェアを開発した. これによってユーザの行動履歴の解析, 検索が容易になり, ユーザやタスクのモデルの学習結果が向上することが期待できる.

また今回の人工知能学会全国大会において, 実際にユーザの行動履歴を収集する実験を行う. 大会中に提供される対話的な情報サービスを利用することにより, 大会中の聴講や関連する情報検索などの行動履歴が収集され, また一部のデータは適切に情報サービスを提供するためにも利用する. こうしたフィールドワークを通じて, 行動履歴収集システムの実用性を検証し, さらに得られたデータの解析や評価を進めることで, 実際のユーザ支援サービスのためのユーザ行動解析システムとして発展させることを考えている. また収集した行動履歴データによるユーザやタスクモデルの学習についても検討を行う予定である.

参考文献

- [1] 本村, 吉田, 藤本: “DSIU(Decision Support for Internet Users) のための Generative User Model”, 人工知能学会全国大会, (2000).
- [2] 本村, 佐藤: “確率的タスクモデリングによるインタラクティブシステムの制御と学習”, 人工知能学会全国大会, (2002).
- [3] 西村拓一, 伊藤日出男, 山本吉伸, 中島秀之: “無電源小型通信端末を用いた位置に基づく状況支援システム”, 情報処理学会研究会報告, 2002-ICII-2, pp.1-6 (2002).
- [4] 松尾, 友部, 橋田, 石塚: “Web からの人間関係ネットワークの抽出と情報支援”, 人工知能学会全国大会 1F1-02, (2003).
- [5] 西村, 橋田, 中島: “イベント空間支援プロジェクト”, 人工知能学会全国大会, 3E1-01, (2003)
- [6] 佐藤, 森, IPA 未踏ソフト 2001 年度成果報告, 人間行動情報蓄積システム, <http://www.ics.t.u-tokyo.ac.jp/ipa/ipa2001/publication/satomori.html>, (2001).