確率ネットワークによるユーザモデル構築システム

User Model Construction System using Probabilistic Networks

本村　陽一[1)]　　原 功[2)]
Yoichi MOTOMURA　Isao HARA

1) 産業技術総合研究所　情報処理研究部門　メディアインタラクショングループ
(〒305-0044 つくば市梅園 1-1-1 中央第二 E-mail: y.motomura@aist.go.jp)

2) 産業技術総合研究所　情報処理研究部門　メディアインタラクショングループ
(〒305-0044 つくば市梅園 1-1-1 中央第二 E-mail: Isao-Hara@aist.go.jp)

ABSTRACT. We developed a user model construction system based on probabilistic networks (Bayesian networks). This system can extract conditional dependencies among variables in SQL databases with conditional probability learning and network structure searching. From a large database of user's feedback, we can construct user models that can predict user's preference. The constructed user models are utilized from other application softwares by sending queries and getting results through TCP/IP connections. We aim to make many information systems more intelligent to react users' intention even from imcomplete observations.

1　背景

最近の情報システムのニーズの高まりと爆発的な普及に伴い、誰もがシステムのメカニズムを深く理解しなくても自然な操作で簡単に望みの結果を得られることが望まれている。そのためにはシステムにとっては完全ではないかも知れないがユーザにとっては自然な入力からシステムがユーザが望んだ通りの動作を選択して、ユーザにとって正しい結果を求めることが必要になる。

ところがこれまでのシステムはあくまでシステムにとって正しい入力が与えられた時に、システムにとって正しい結果を出力するため、ユーザ側でシステムの各機能や動作メカニズムを理解し完全に正しく操作しなければならないという弊害が存在する。また、画一的な利用法だけでなく様々な環境で幅広いユーザに利用されるようになるにつれ、システム設計者が全ての起こり得る状況を事前に考慮することが困難になるという問題もある。一方で WWW を通じた電子商取引の普及に伴い、各個人の個性や嗜好性にシステムが対応するパーソナライゼーションや CRM(Customer Relationship Management) の重要性も強く認識されてきている。

こうした要請に応えるためには、情報システムがユーザの意図や嗜好性などを推定する能力を持つことが必要である。ユーザに関する予測を行うために内部的にシミュレートを行うモデルがユーザモデルである。このユーザモデルを統計データからの学習によって構築することができれば、システム自らが使用されている状況に適応していくことも可能になる。

そこで本プロジェクトではこのユーザモデルを統計データから構築し、さらにこのユーザモデルを様々な他のシステムが利用できるようにするためのシステム開発を行った。とくに状況に応じて非決定的な挙動を示す人間の不確定性をモデル化するために確率ネットワークを用い、本格的な SQL データベース内の統計データからの学習によってユーザモデルを構築することが本提案の特長である。

2　目的

今回我々が目的とするのは、多くのプログラム開発者が利用できるようなユーザモデルサーバー、ユーザーモデル構築システムである。つまり、ある特定のアプリケーションを開発対象にするのではなく、多くのアプリケーションソフトがユーザモデルを簡単に利用できるような汎用の枠組を提供する。これにより、多くのシステム開発者を支援し、ユーザに合わせて動作する使いやすいアプリケーションシステムが社会に広く普及することを目指し、日本のソフトウェア技術全体のレベルアップに貢献することを狙う。

3　確率ネットによるユーザモデル

3.1　ユーザモデル

従来の情報システムの多くはユーザ自身がコマンドとシステムの挙動 (システムモデル) を習得し、状況に応じて正しく操作することが要求される。一方、シス

テムの側がユーザに関するモデル (ユーザモデル) を持ちユーザの意図や要求を正しく推定できれば、ユーザの方はごく自然に振舞うだけで簡単にシステムを利用することができる。

ただしユーザの意図の推定は完全に観測できるものではないため、不確定性をうまく取り扱い個人の特徴づけをタスクやシステムに依存しない一般的な表現でモデル化することが重要である。そこでここでは確率的な枠組によるモデル化を行う。

ユーザモデルは、ユーザの嗜好性や意図などの通常簡単に観測することのできない要素 (隠れ変数) を、他の観測の容易な要素 (観測変数) から予測するものになる。このユーザモデルを構築するためには、どの要素 (変数) に注目して表現するか、モデルをどのような構造で実現するかなどを統計データに基づいて解析し発見することが必要になる。そこで確率変数の依存関係をネットワーク状にモデル化した確率ネットワークを用いて、隠れた要素を予測するモデルを構築する。

3.2 確率ネットワーク (ベイジアンネット)

ノイズや不確定な要因を含む不完全な観測情報を取り扱うことが必要である時、確率ネットワークを使って対象をモデル化することで知りたい変数の確率分布を推定し、起こり得る各状態の確率 (確信度) を評価する方法がある。特に問題対象の背景にある複雑な依存関係を表すためにグラフ構造を利用し、依存関係のある変数の間を向きを持ったリンクで結び、リンクをたどったパスが循環しないような非循環有向グラフで表される確率モデルがベイジアンネット [1, 2, 3, 4] である。

まず、確率的な変数 X,Y の間の依存関係を条件付確率 $P(Y|X)$ で表す。これは X のとる値に応じて、Y の分布が影響を受け、その依存関係の定量的関係が条件付確率分布 $P(Y|X)$ で定められるということである。ベイジアンネットはこうした変数間の依存関係を全て結んだネットワーク構造を使って、ある観測可能な変数の値が得られた時に、他の観測不可能な変数の予測を行うことができる。

確率変数 X_i, X_j の間の条件付依存性をベイジアンネットワークでは向きのついたリンクによって $X_i \rightarrow X_j$ と表し、X_i を親ノード、X_j は子ノードと呼ぶ。親ノードが複数あるとき子ノード X_j の親ノードの集合を $\pi(X_j) = \{X_1, \cdots, X_i\}$ と書くことにする。

この時の変数 X_j の値が親ノードの変数の値によって影響をうけるが、それが非決定的、つまり親ノードの値だけにはよらない不確実性がある時、この関係を子ノードの変数 X_j について親ノードの値を条件とする条件付確率,

$$P(X_j \mid \pi(X_j)) \qquad (1)$$

で表すことができる。この一つの子ノードについての関係はベイジアンネットの中で X_j を子ノード, $\pi(X_j)$ を親ノード群とする局所的な木になっている。[1]

確率変数が離散的な場合、条件付確率は全ての状態における確率値を並べた表、CPT(Conditional Probability Table) によって過不足なく表すことができる。例えば親ノードがある状態 $\pi(X_j) = \boldsymbol{y}$ ($\boldsymbol{y}$ は親ノード群の各値で構成したベクトル) のもとでの n 通りの離散状態を持つ変数 X_j の条件付確率分布を $p(X_j = x_1|\boldsymbol{y}), \cdots, p(X_j = x_n|\boldsymbol{y})$ とする (ただし $\sum_{i=1}^{n} p(x_i|\boldsymbol{y}) = 1.0$)。これを行として、親ノードがとりえる全ての可能な状態 $\pi(X_j) = \boldsymbol{y_1}, \cdots, \boldsymbol{y_m}$ について列を構成した表 1 が X_j にとっての CPT、$P(X_j|\pi(X_j))$ である。

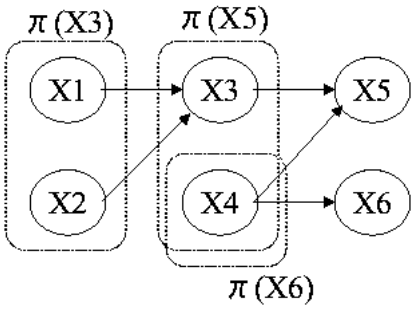

図 1: ベイジアンネットワークの例

表 1: 条件付確率表 (CPT)

$p(X_j = x_1\|\pi(X_j) = \boldsymbol{y_1})$	$\cdots$	$p(X_j = x_n\|\pi(X_j) = \boldsymbol{y_1})$
:	$\ddots$	:
$p(X_j = x_1\|\pi(X_j) = \boldsymbol{y_m})$	$\cdots$	$p(X_j = x_n\|\pi(X_j) = \boldsymbol{y_m})$

さて、ベイジアンネットを実際に構築する手順は以下の通りになる。

- モデルで使用する確率変数, X_j を決定しノードを作成する。
- 変数間の依存関係にしたがって、親ノードから子ノードにリンクを張っていく ($\pi(X_j)$ の決定)。
- 変数間の依存関係を定量的に表す条件付確率表 (CPT, $P(X_j|\pi(X_j))$) を決定する。

3.3 ユーザモデルのための確率ネット

ユーザモデルのためにベイジアンネットワークを適用することは国際的には最近かなり精力的に研究が進められている。例えばマイクロソフトでは、ユーザにヘルプ情報を与えるためにベイジアンネットに基づくユーザモデルを使用し、マウスの移動や直前の操作からユーザが操作方法について迷っているかどうかを予測している。[5]。しかし、そこではユーザは均質の性質を持つものとして扱われており、それぞれのユーザの個性や特徴が十分に反映されているとは言えない。[2]

[1] この局所木はデータマイニングでよく用いられる決定木のもっとも大きい場合になっている。つまりベイジアンネットにより表される確率モデルは決定木をその部分クラスとして含むより一般的なものと言える。

[2] そのためか、これまで MS のオフィスアシスタントを煩わしいと感じられたユーザも少なくないであろう。

一方、我々が目指すのは個々のユーザに合わせて個別に適応できるものである。そこで個人やアプリケーションに合わせて動的にモデルを構築する必要がある。そのために留意すべきポイントをあげてみる。

- 実際にユーザモデルを使用する状況で収集した大量のユーザ履歴データから統計的学習を行ってモデルが構築できるように本格的な SQL データベースと連携できること。
- ユーザの挙動や意図などのあらかじめ明示的に定義することが難しい対象をモデル化するために、柔軟に変数 (ノード) の定義を変えながら対話的にモデル構築ができること。
- 変数間の依存関係を示すネットワーク構造もあらかじめ明示的に定義することが容易でないために、自動的に最適な構造を探索できること。(この時複数のモデルの候補の中から、客観的な基準 (情報量基準) によって決定する。)

これらを実現できるようにユーザモデル構築システムを開発した。

4 ユーザモデル構築システム

本システム (図 2) の開発は JAVA 言語で行い JDBC による SQL データベース接続機能を利用することでユーザの印象、行動履歴などに関する大量のデータからでもユーザモデルが構築できる。またユーザモデル、確率ネットワークに関する研究はマイクロソフトを始めとして、国際的に研究開発の速度が速く、逐次高度な学習・推論アルゴリズムを開発し、評価することも重要である。そこで JAVA のリフレクションによって一部のクラスだけをコンパイル、追加することで簡単にシステムを拡張できる機能を実装し、多くの研究者が自身のアイデアを大規模なデータに対して簡単に実験でき、国内のユーザモデルや確率ネットワーク研究を促進することも視野に入れている。

本システムでは、データベースに格納された統計データを検索し、モデルとの適合性を確認しながら変数、ネットワーク構造、条件付確率というモデルの各要素を決定していく。ネットワーク構造の決定は尤度、情報量基準 MDL, AIC などのモデル選択基準を用い、これは利用者が自由に選ぶことができる。以降ではユーザモデルの構築手順に合わせて本システムの内容を説明する。

4.1 ユーザモデル構築の概要

まずデータベースに格納された統計データの中から適切なデータセットを取り出す。ユーザモデル構築システムではこの部分的なデータセットに基づいて依存関係を評価し、ユーザモデルを構築していく。本システムで構築するユーザモデルとは、先に説明したような依存関係のある確率変数を抽出し、その間にリンクを張って構成したネットワークと、この変数の関係を定量的に表す条件付確率パラメータである。そこで、与えられたデータにもっとも良く適合するような変数の集合と、その変数を結合したネットワーク構造、そして変数間の条件付確率を決定していく[3]。

ここでネットワーク構造の選択基準はとくに重要な研究課題であり、これまで、条件付依存性の強さやデータへのフィッティングの度合、モデルの複雑さなどの基準 (情報量基準) として様々なものが考えられている。そこで本システムはこうした様々な情報量基準を自由に選んだり、使用者が独自に追加できるようにする。デフォルトでは AIC と MDL を利用したものを用意し選択でき、またユーザが独自に開発したモジュールも簡単に追加して使用できるようになっている。

変数の生成、割り当てと条件付確率の計算は SQL データベースと接続し、対話的に操作できるため非常に効率的である。条件付確率パラメータはデータベース中の頻度から自動的に計算しシステム内のテーブルに格納する。実際のデータベースの中には全ての組み合わせのサンプルが存在しないことがあり、テーブルが完全に埋まらない (確率パラメータが得られない) という問題が生じる。そこで、本システムは得られているデータだけから欠けているデータを補完する、条件付確率の学習機能を導入した。デフォルトではニューラルネットワークが用意されているが、JAVA の継承、リフレクションにより、使用者が独自に開発した他の学習モジュールを簡単に追加することができる。

システムは JAVA(JDK1.3) により実装されており、Swing を利用した豊富な GUI(グラフィックインターフェース) によるモデルの可視化、JDBC により広く一般に使われている (Postgres や Oracle などの) 主要な各種データベースとの接続性の良さ、オブジェクト指向アーキテクチャによる拡張性の良さなどが特長である。

4.2 データベース操作機能

とくにこのシステム独自の特徴として JDBC により、よく使われる SQL データベースと連携することで、従来はメモリ消費が激しく実行が難しくなるような大量のデータに対しても SQL 検索コマンドを用いた高度な操作が可能である。また、データとしては陽に格納されていないような変数 (例えばある 2 つの時刻の差分など) でも、SQL データベースの演算操作によって実現できればベイジアンネットの変数としてその場で利用することができる。

具体的な操作方法を見てみよう。まず SQL データベースと接続すると、その中のテーブルの一覧が表示され、さらにテーブルを選ぶと、全ての項目 (ユーザの性別や年齢などの属性やアンケートの回答) が表示される。そこである項目を選択すると、ベイジアンネットのノードが生成され、ウィンドウに表示される。このようにしてユーザモデルとして必要な項目を全て選択する。次に予測したい項目 (例えば「ある対象をどの位好きか」) を子ノードとして、関連する他のノードを親ノードとしてリンクを張る。そこで条件付確率の計算を実行すると、この依存関係についてのデータベース内のデータの頻度をカウントし、さらにそれを

[3] これはデータベースから知識発見を行うデータマイニングと同様の操作であり、実際ベイジアンネットは決定木の他にもデータマイニングでよく使われる相関マッチング、ニューラルネットなどの多くのモデルの一般形として理解することもできるため、包括的なモデル化が可能である。

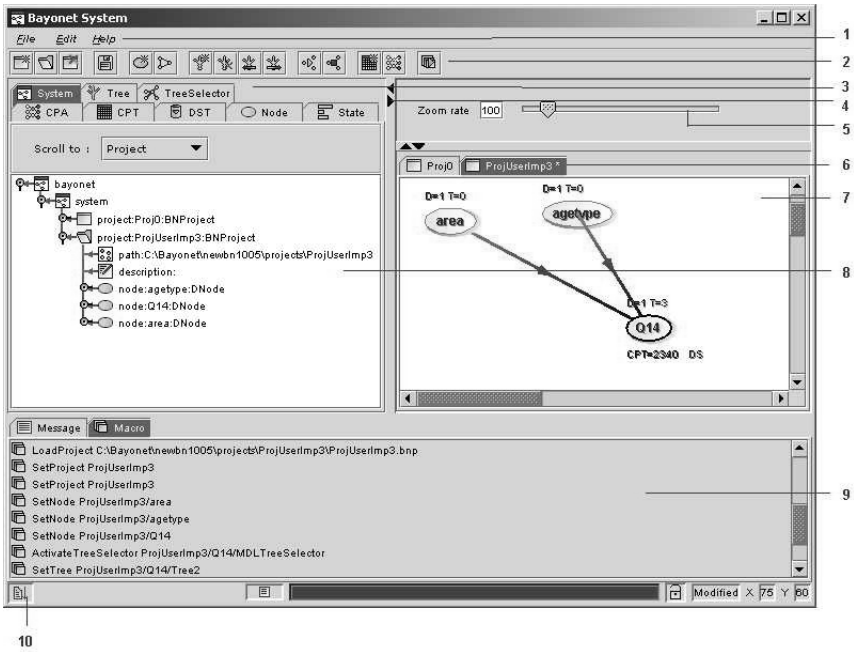

図 2: 開発したユーザモデル構築システム

正規化した条件付確率が自動的に計算される。同時に現在のモデルに含まれる部分モデルについての計算も同時に行う。

4.3 モデル選択機能

こうして得られたデータの頻度と、条件付確率表からモデルの平均対数尤度や変数間の相互情報量を計算する。平均対数尤度、つまり予測精度が高くなればこの親ノードと子ノードの条件付依存関係が強いということを表す。そこで、これらの内部で計算した定量的な指標からシステム内の全モデルから最適なものを探索する。これを行うために、本システムは MDL や AIC などのいくつかのモデル選択基準を内蔵する他、ユーザが独自のモデル選択アルゴリズムを追加できるようになっている。

4.4 欠測データからの条件付確率学習機能

変数が離散的でかつ全ての組合せを含んでいる完全データの場合には、先に述べたような手順でデータベース中の頻度を計算し、全ての条件付確率値を求めることができる。しかしユーザのアンケートデータはしばしば不完全であり欠測データを含むため、データの頻度だけでは全ての条件付確率値が得られないという問題が生じる。

この場合は周辺のデータによって欠測データに関する条件付確率を推定することが必要となる。この欠測データの推定には親ノードの値を入力、その時の子ノードの確率値を出力とするようなニューラルネットや回帰モデルなどの学習モデルを利用することを考える。すでにわかっているデータの頻度から求めた確率値を教師信号としてモデルを学習し、これで補完したものを欠測データについての条件付き確率として用いる。これは学習モデルの汎化能力を期待して未学習の条件付確率を近似していることになる。なおニューラルネット以外に回帰モデルやサポートベクターマシンなど、他のモデルを追加することも可能になっている。

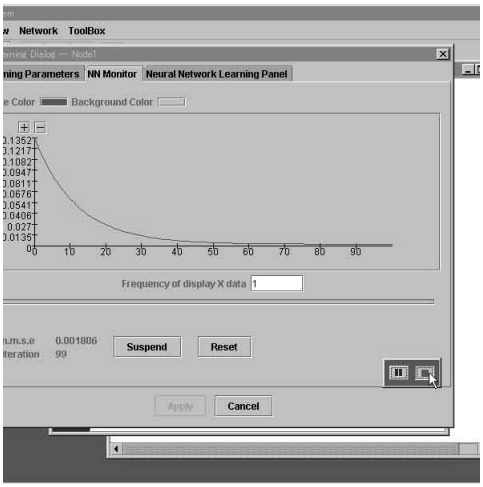

図 3: ニューラルネットによる条件付確率の学習

4.5 システムの拡張性、接続性

本システムは次のような特長により他のプログラムと連携して利用することができる。

- JDBCドライバを持つ主要な各種データベースシステムとの接続
- JAVAのリフレクションによる、各種プログラムモジュールの追加
- XMLや、他のベイジアンネットソフト(Huginなど)との互換ファイル生成機能
- TCP/IPコネクションによる外部プログラムとのインタフェース

これらの特長により様々な問題に対する実用的な大規模なデータベースとの接続が容易になり、また条件付確率の近似のための学習モデルやモデル選択アルゴリズムなどを追加、選択して同一条件での性能評価を行うことも容易になる。したがって、本システムをフリーソフトとして公開しインプリメントに比較的手間のかかるグラフィカルユーザインターフェースやデータ管理部などを統一的に提供することで、各利用者は様々な最新の学習モデルやアルゴリズムの部分だけを実装し、短期間のうちに大規模データベースを利用して評価できるため、ユーザモデル、確率モデル研究の活性化にも寄与できると考えている。

5 ユーザモデル構築システムの評価

5.1 ユーザモデルの例

確率モデルを利用したユーザモデルの例として、次のようなものを考え、これを評価用サンプルとして開発システムのテストを行った。

- U:ユーザの特徴(性別、年齢、職業など)を数次元のベクトルで表す。
- X:画像やホームページなどを対象とし、その属性(画像特徴やキーワードなど)を数次元のベクトルで表す
- E:ユーザがその対象をどのくらい「好き」だと思うかについての印象を段階的に評価したもの(「とても好き」、「好きではない」、「わからない」など)
- ユーザモデル:U、XからEを確率的に予測する。

例:

$$P(E=\text{"とても好き"}|U,X)=70\% \quad (2)$$
$$P(E=\text{"やや好き"}|U,X)=15\% \quad (3)$$
$$P(E=\text{"わからない"}|U,X)=10\% \quad (4)$$
$$P(E=\text{"好きではない"}|U,X)=5\% \quad (5)$$

多数の披検者に様々な対象Xを評価してもらい、その対象に対してどのような印象を持ったかについてのアンケートデータを集め、回答の頻度から条件付確率パラメータを獲得する。こうして集めたデータから情報量の高い属性、特徴Uを抽出し、より良くEを予測できるようにモデルを構築していく。

5.2 アンケートデータによる実行例

あるWWWページを見た時の印象(「その対象をどの位好ましいと思うか?」)について10代から60代までの様々な職種の男女550人にアンケートを取った結果を用いてそのWWWページに対する好みを予測するユーザモデルを構築する例を示す。

まず、新しいプロジェクトを開いてから、DataSetPanel(DST)のタブを開きデータベースと接続する。アンケートを格納したテーブルを選択するとその中の項目が表示されるので、「その対象をどの位好きか?」というアンケート項目を子ノード、職業や性別などのユーザ属性や他のアンケート項目を親ノードとして選びノードを作成していく(図4)。

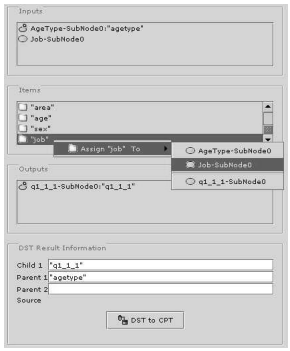

図4: アンケートデータの項目に対応したノードの生成

可能性のある親ノードを全て選んだら、それをグループとして選び、子ノードへのリンクを張る。そこで"DSTtoCPT"アイコンを押すとデータベース中のアンケートの頻度から条件付確率が計算される。この状態で"DecomposeTree"を実行すると、現在の木構造に含まれる、全ての部分木をモデルの候補として内部的に生成する。次にTreeSelectorパネルへ移り、モデル選択アルゴリズムを選ぶと、内部に保持した全ての候補から最適なモデルを一つ決定する(図5)。

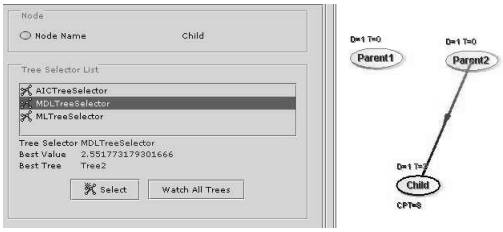

図5: 最適なモデルの選択

以上の操作を繰り返し、ネットワークを構築していくことでユーザモデルを作成する。完成したユーザモデルに対して観測可能な変数の値を代入すると、予測したい対象に関する確率値を得ることができる。これを外部のアプリケーションが利用することでユーザの意図や好みを予測し、それに応じて適切な動作を実行することが可能になる。

6　まとめ

ベイジアンネットに基づいたユーザモデル構築システムを開発した。とくにこのシステム独自の特徴として、SQL データベースと連携することでデータ量に対してスケーラブルであり、SQL 検索コマンドを用いた高度な検索により適切な変数選択が可能になる。また局所的に複数の木を作成しておき、その中から情報量基準にしたがって最適な木構造を自動的に探索することでグラフ構造を決定していく仕組みを導入した。さらにニューラルネットを用いて学習することで欠測データがある場合やデータ数が十分でない場合でも条件付確率値を近似することを可能にした。また本システムでは、ニューラルネットの他の学習モデル、モデル選択アルゴリズムなどを追加、拡張可能にしており、利用者が新たなモデル、アルゴリズムを開発し、実験評価することも容易になっている。これらの特長はこれまでに存在するベイジアンネットソフトウェアには見られないものであり、国際的にも WWW を通じて多くの研究者、開発者から問い合わせがあり、ダウンロードされている。

本システムを WWW や CD-R にて一般に無料で公開することで多くのシステム開発者がユーザモデル、確率ネットワークを利用できるようになる。今回のプロジェクトを通じて得られた成果に関しては WWW ページ http://www.aist.go.jp/ETL/~motomura/IPA/ を開設し、成果の普及をはかっている他、これまでに数回のデモ、CD ー R の配布、チュートリアルの開催, 国際会議での発表などを行った [6, 7, 8, 9]。

今後も次のような波及効果を期待し、本プロジェクトの成果普及に努める。

- ユーザモデルの構築ツールと幅広い実用システムへの応用例を示すことで誰でも安心して使えるユーザフレンドリな高度で複雑な情報機器の開発に貢献する。
- ユーザモデルの構築のために整備されるデータベースと構築した各種のユーザモデルが共有、標準化できることによって当該分野の研究促進と成果の効果的な普及をはかる。
- ユーザモデル利用技術が広く普及することで、当該分野における人間に優しいソフトウェア技術の向上に寄与する。
- 様々なユーザが情報システムを操作する際の挙動を解析することにより、認知科学的側面からの研究にも寄与する。

様々なアプリケーション、ドメインごとにデータベースを整備し、ユーザモデル構築のための解析を集約的に進めることも重要である。またデータに基づくユーザモデル構築のためにはモデルが複雑になるにつれて、必要とされるデータ量も膨大になる。そこで今後の課題としては、データ量が不十分な場合でも背景知識などを補うことでモデルを構築する技術開発がとくに重要であり、そのために一階述語論理表現から確率ネットワーク表現へ変換する方法などの研究を進める。

7　謝辞

プロジェクトの実施にあたり、プロジェクトマネーシャーの京都大学教授上林弥彦教授には終始ご指導、貴重な機会を与えていただいたことを深く感謝いたします。また、プロジェクト管理組織の京都高度技術研究所、三好様、杉本様、平家様にはプロジェクトの実施を支援していただきました。アンケートデータの収集にあたっては九工大吉田氏に尽力していただきました。そのほか、シリコンバレーの調査や、京都での報告会などの機会を通じて、ご意見をいただきました皆様、未踏ソフトウェア創造事業を担当された IPA の関係各位にこの場を借りて感謝いたします。

8　参加企業及び機関

財団法人京都高度技術研究所

参考文献

[1] Castillo, E., Gutierrez, J., and Hadi, A.: *Expert Systems and Probabilistic Network Models*, Springer-Verlag (1997).

[2] 石塚満 (訳)：15 章: 確率的推論システム, 古川康一監訳, エージェントアプローチ人工知能, pp. 439–473, 共立出版 (1997).

[3] 本村陽一, ベイジアンネットワーク, 電子情報通信学会誌, Vol.83,No.8, pp.645–646 (2000).

[4] 本村陽一, 佐藤泰介, ベイジアンネットワーク–不確定性のモデリング技術–, 人工知能学会論文誌,*vol.15, No.4*, pp. 575–582 (2000).

[5] Horvitz, E., Breese, J., Heckerman, D., Hovel, D., and Rommelse, K.: The Lumiere Project: Bayesian User Modeling for Inferring the Goals and Needs of Software Users, *Proc. of the 14th Conference on Uncertainty in Artificial Intelligence*, pp. 256–265 (1998).

[6] ベイジアンネット　チュートリアル BN2001, 人工知能学会基礎論研究会主催, http://www.aist.go.jp/ETL/~motomura/bn2001/ (2001).

[7] 本村陽一: データベースと連携したベイジアンネット構築システム, 第 4 回情報論的学習理論ワークショップ IBIS2001 予稿集, pp.223–226 (2001).

[8] Motomura,Y., Yoshida,K. and Fujimoto,K., Generative User Models for Adaptive Information Retrieval, proc. of IEEE System, Man and Cybernetics, pp. 665–670 (2000).

[9] Motomura,Y., et.al., Task, Situation and User Models for Personal Robots, IJCAI 2001 workshop on Reasoning with Uncertainty in Robotics, pp.51–56 (2001).