1

KDDカップ2000主催者報告書：本質をつかむ

翻訳：鈴木英之進（横浜国立大学工学部電子情報工学科）
suzuki@dnj.ynu.ac.jp

KDD-Cup 2000 Organizers' Report: Peeling the Onion

Ron Kohavi (Blue Martini Software)
ronnyk@bluemartini.com
Brian Frasca (Blue Martini Software)
brianf@bluemartini.com
Zijian Zheng (Blue Martini Software)
zijian@bluemartini.com
Carla E. Brodley (Purdue Univ.)
brodley@ecn.purdue.edu
Llew Mason (Blue Martini Software)
lmason@bluemartini.com

First appeared in SIGKDD Explorations, Volume 2, Issue 2
(http://www.acm.org/sigkdd/explorations/).

KDDカップは1997年から毎年行われているコンテストである．第4回目となるKDDカップ2000は，前主催者の提案に従い複数個の問題を出した．今回初めて，KDDカップは予測問題に加えて洞察問題を出した．

出題領域はeコマースであり，これはデータマイニングの成功に必要な要素[1]をすべて含んでいるため魅力的な領域である．必要な要素とは，(i) 多数の属性，(ii) 多数のレコード，(iii) 管理されたデータ収集（電子的な収集），(iv) 評価可能な結果と明示的な投資リターン，および (v) 行動（Webサイトの修正やクロスセリングの提供など）が容易に起こせる領域である．ブルーマティーニソフトウェア社は，顧客インタラクションシステムを通してデータを提供してくれるようにクライアント数社に要請し，ガゼルドットコム社というレッグウェアとレッグケアの小売り業者である小さなドットコム会社が機密資料を適切に削除することを条件にボランティアとしてデータを提供することに同意した．

データを分析しガゼルドットコム社とブルーマティーニソフトウェア社の小売り専門家と相談することで，5個の問題を定義した．これらのうち2個は予測問題であり，残りの3個は洞察問題である．各参加グループには入手したデータの一部だけ（最初の約2カ月間）を渡し，評価のためのテスト集合（3カ月目）は非公開とした．

問題をより現実的にするために，ガゼルドットコム社に実情を説明してもらい，各参加グループが同社のマーケティングカレンダーを見られるようにした．カレンダー上のイベント（たとえばTVコマーシャル）は，時期に応じて変化するWebサイトへの訪問者数をよく説明している．

データは，オリジナル版と集約版の2種類の形式で提供した．オリジナルデータはページリクエストレベルで収集されたが，コンテストの問題はセッションレベルや顧客レベルで出されている．既存のほとんどのツールは集約能力が不十分であるため，ブルーマティーニ社の顧客インタラクションシステムを使ってセッションレベルや顧客レベルでの振る舞いを要約した集約データを生成した．データと集約方法については「使用データ」の章で詳しく述べる．

洞察問題はブルーマティーニ社の小売り専門家の助言を受けて評価した．我々は「コンテストの問題と評価基準」の章で説明する標準的な採点方法を設定した．「投稿の統計値」の章では投稿結果の各種統計値を示すが，投稿結果を評価した際に多くの観察が不十分であり諸問題点の深い理解につながらないパターンを含むことに気づいた．我々はデータマイニング調査を行う際に本質をつかむことの重要性を強調したい．「結果と洞察」の章では結果と洞察を説明する．最後に得られた教訓を示し結論とする．コンテストの優勝者からのレポートはWebサイト[2]や文献8）に掲載されている．

コンテスト問題の背景

コンテスト問題をよりよく理解するため，ガゼルドットコム社のWebショップについて実情を数点述べる．

• ホームページには70個以上の画像があるため，モデムを使用するユーザにとってダウンロードがきわめて遅い．

• 多くのドットコム会社と同じように，ガゼルドットコム社の当初の目的は短期的には損失を出しても多くの顧客を引きつけることだった．同社は多数の販売促進キャンペーンを行い，これらのキャンペーンはWebサイトのトラフィックや顧客の種類に影響するのでマイニングにおいて重要である．重要なキャンペーンは次のとおりである．

①FREE: 3/20から4/30の間，送料無料（3.95$の価値）．（通常は支払い額が40$を超えた場合に送料無料．）

②MARCH1: 3/1から4/1の間，10$割引．

③FRIEND: 2/29から4/30の間，10$割引．

④FREEBAG: 3/30から4/30の間，カバンを無料提供．

• 2/28，ガゼルドットコム社はゴールデンアワー時に放映される人気コメディショー番組Ally McBealでTVコマーシャルを放映した．

• 2/26，ガゼルドットコム社は申込書の形式を大きく変更した．このため収集される顧客属性がこの日を境に変化した．

コンテストの問題と評価基準

KDDカップ2000では5個の問題が出題された．うち2問は客観的な評価基準がある標準的な予測問題であり，残りの3問は主観的な「洞察」問題である．

問題1：Webページ閲覧履歴から，訪問者が別のページを見るかWebサイトを去るかを予測せよ．

出題の動機は，訪問者がWebサイトを去りそうであるかを知ることは，表示すべき最良のページ（たとえば訪問者を去らせないための特別な販売促進）を決定する材料となるというアイディアである．この問題の評価基準は単純にテスト集合における正しい予測数である．優勝者は正答率が最も高い参加グループである．

問題2：セッション初めのWebページ閲覧履歴から，訪問者がその後見る製品のブランド（ヘインズ，ダナキャラン，アメリカンエッセンシャルズ，あるいはその他）を予測せよ．

この問題の動機は，特定のブランドが載っているページへのハイパーリンクを自動的に張ることによりナビゲーションを容易とする課題である．問題を過度に難しくしないために，予測対象を最もよく売れている3ブランドのうちの1つかその他（3ブランドのいずれも残りのセッションで見ないと定義される）の2種類とした．評価基準は次の方法で計算する重みつき予測得点である．

2点：3ブランドの1つを予想し，残りのセッションにその予想ブランドのページが含まれる場合．

1点：その他を予想し，残りのセッションに3ブランドのページがまったく含まれない場合．

0点：他の場合．

最高得点者を優勝者とした．

残りの3問については，参加グループに産業界のユーザが理解でき有用と認める文章とグラフを提出することを求めた．文章は1,000語以下，グラフは10個以下とした．

問題3：一定期間の購入履歴から，Webサイト上の1回の注文で平均12$以上支払う訪問者を特徴付けよ．

この問題の動機は，価値が比較的高い顧客についての洞察がマーケティングの方向や製品の選定などに有用となることである．

問題4：問題1と同じであるが，正確な予測よりも洞察を重視した．

問題5：問題2と同じであるが，正確な予測よりも洞察を重視した．

問題3，4，5については単純な客観的評価基準がないため，ガゼルドットコム社とブルーマティーニ社の小売り専門家に相談した．この結果，次の方法で評価を形式化した．まずすべての重要な洞察を集めてそれらを重み付けし，次に発見された洞察，レポートの正確さ，およびレポートのプレゼンテーション（産業界のユーザがターゲットであることを念頭において）に基づいた得点を組み合わせた評価基準を考案した．参加グループを評価するために実際に用いた得点の計算式は次のとおりである．

$$Score=3P+3C+\sum_{i=1}^{N} w_i I_i$$

ただし$\boldsymbol{P}$（0−10）は参加グループのプレゼンテーション得点，$\boldsymbol{C}$（0−10）は参加グループの正確さ得点，そして各洞察iについてw_iは割り当てられた重みであり$\boldsymbol{I}_i$（0−2）は参加グループの得点である．洞察の数と重みは問題によって異なる．プレゼンテーション得点は参加グループが提出した投稿のプレゼンテーションの良さを反映し，たとえば次の要素を評価した．

• 投稿の読みやすさと理解しやすさ．

• 産業界のユーザたちがグラフ，表，および図を理解できたか．

• 重要な情報を把握する労力が必要か，あるいは無関係な情報が大量にあるか．

正確さ得点は参加グループの論点が正確であるかという点と，各論点を支持する十分なデータが示されているかという点に基づいてつけた．

各問題に対して，各参加グループが提出したすべての洞察に基づいて完全な洞察リストを定義した．これらの洞察を，ガゼルドットコム社とブルーマティーニ社の小売り専門家との議論に基づき，産業界のユーザにとっての興味深さを反映するように重みづけした．多くの洞察は単純に，より基本的な洞察と関連するので重みを低くした（重み0という場合もあった）．各参加グループに対して，各洞察に0点（その洞察を発見しなかった場合），1点（その洞察を部分的に発見した場合），あるいは2点（その洞察を完全に記述した場合）のいずれかの得点を与えた．洞察は多数あるため（質問3と4に対して各30以上），本稿ではなくKDDカップ2000のホームページ[3]に示す．

使用データ

本章ではWebストアで収集されたデータ，データウェアハウスから初期スタースキーマを構成する方法，用いたデータクレンジング／隠匿方法の種類，および適用したデータ変換の種類を説明する．最後にKDDカップで用いた最終スキーマとデータ形式を要約する．

初期データ収集

2000年1月30日，ガゼルドットコム社はブルーマティーニ社の顧客インタラクションシステム(CIS)を使って，社員の友人や家族を顧客として試験的に起業した．Webストアでは，ブルーマティーニ社のアーキテクチャによるアプリケーションサーバがJavaで書かれたテンプレートに基づいてWebページを生成する．このアーキテクチャでは，主に顧客のトランザクションとクリックストリームに関するログをアプリケーションサーバ層で記録する．アプリケーションサーバがコンテンツ（例：画像，製品，および記事）を生成するため，サーバで供給するコンテンツは詳しく分かっている．このことは，精算の際などにコンテンツが動的に生成される場合や通信用に暗号化される場合にも当てはまる．Webログデータは必要ない．アプリケーションサーバは，ユーザのセッションを追跡するためにクッキーを使用する（クッキーがない場合にはURLを符号化して用いる）．よって通常のWebログに必要な，クリックストリームをセッションにまとめる作業は必要ない．アプリケーションサーバはログイン機構やクッキーを使用するユーザも追跡するので，個々のページ閲覧を特定の訪問者に関連付けることは容易である．

ブルーマティーニ社のアプリケーションサーバによって収集されたデータのうち，次の3種類はこのKDDカップに関連する．

- 顧客情報：顧客ID，登録情報，および登録時のアンケート回答．
- 2レベルの注文情報：1）日時，割引，税，支払総額，支払方法，発送，状態，およびセッションIDを含む注文ヘッダ；2）数量，価格，製品，日時，分類，および状態を含む注文行．
- 2レベルのクリックストリーム情報：1）開始終了日時，クッキー，ブラウザ，参照元Webサイト，訪問回数，およびユーザエージェントを含むセッション；2）日時，シーケンス番号，URL，処理時間，製品，および分類を含むページ閲覧．

各顧客は通常複数のセッションに関係する．各セッションは複数のページ閲覧と複数の注文に関係する．各注文は複数の注文行から構成される場合がある．各注文行は1個以上の単一製品に関する購買履歴を表す．

スタースキーマの生成

ブルーマティーニ社のアプリケーションサーバに含まれるデータ収集機能は，オンライントランザクション処理（OLTP）システム内で実装されている．OLTPシステムは多数の小さな更新や短い問合せを効率的に処理するように設計されている．これはeコマースビジネスを運営する上ではきわめて重要であるが，分析では通常いくつかの大規模なテーブルや産業界のユーザが理解できるスタースキーマデザイン[6),7)]を何回か完全にスキャンする必要があるため，分析には適していない☆1．

データマイニングのためには，次元モデリング技術を使ってデータウェアハウスを構築する必要がある．データウェアハウスの設計も，OLTPからデータウェアハウスシステムへのデータ変換も両方とも大変複雑で時間がかかるタスクである．ブルーマティーニ社のアーキテクチャはテーブル，行，そしてそれらの関係についてのメタデータを含むため，OLTPシステムから自動的にデータウェアハウスを構築できる[5)]．

KDDカップ用のデータを準備する際に，我々はアクシオムから発表されているデータをスキーマに統合した．この結果分析用の顧客情報に，性別，職業，年齢，結婚状態，推定年収，および家の資産価値など50以上の新しい属性が加わった．

☆1 スタースキーマとはデータベーステーブルのジョインであり，中央に1つのファクトテーブルがありそれが次元と呼ばれる複数個のテーブルとジョインされている．

KDDカップのデータを生成するために，クリックストリームスターと注文行スターの2種類のスタースキーマが使用された．クリックストリームスターはクリックストリームのファクトテーブルと6次元のテーブル，すなわち顧客プロファイル，アクシオム，Webセッション，製品，分類，およびコンテンツから構成される．注文行スターも注文行のファクトテーブルと6次元のテーブル，すなわち顧客プロファイル，アクシオム，注文ヘッダ，製品，分類，および販売促進から構成される．

データクレンジング／隠匿

顧客のプライバシーを保護するため，ログイン名，パスワード，クレジットカード，顧客名，およびセッションIPアドレスなどの個人情報を削除した．製品単位コストなどの利益に関する情報も削除した．このデータをマイニングする（KDDカップの問題を解く）のに重要だと考えた属性については，データを変換した．たとえば，電子メール属性の値は，COM，EDU，ORG，およびGOVなどドメイン接尾辞だけを残すように変換した．これに加えて電子メールアドレスのGazelle.comをgazelle.comの接尾辞で保存した．すべての会社名はCOMPANYと1つの数字に変換した．この結果，同一会社に所属する人は分かるが会社名は分からなくなった．セッションクッキーIDを符号化し，各クッキーIDを元とは異なる値にした．ただし同一クッキーに関するセッションは分かる．

データクレンジングは通常KDDプロセスの一部分である．我々は以下に述べる3つの理由により自分たちで初歩的なデータクレンジングを行った．最初に本当のデータマイニングプロジェクトと違い，KDDカップの各参加グループは領域専門家に相談できない．次にデータを公開する前にデータを隠匿しなければならない．最後にこの初歩的なデータクレンジングを行った後でさえ，出題問題は十分難しい．データを次の手順でクレンジングした．

- 要約レコードを削除した．要約レコードは曜日と時間に関係なく1分間に3回現れ，1カ月間で約125,000セッションを生成する．これらのレコードによりマイニング結果が偏ってしまう．
- テストユーザを削除した．我々は，testが入った顧客名は除く，異なる15人以上のユーザに使われたクレジットカードでの購入は除くなどいくつかの基準を設けた．テストユーザは購買や閲覧の振る舞いが通常とは大変異なることに注意されたい．
- 返品注文と未完注文を削除した．これらは少数だが混乱を招くかもしれない．

セッションID	要求シーケンス	問題1テスト集合	問題2テスト集合	セッション続行
29	1	T	T	F
29	2	T	T	F
29	3	T	F	F
56	1	T	T	T
56	2	T	T	T
56	3	F	F	T
68	1	T	T	F

表-1 セッションの切り取り方法

データ変換

KDDカップでは非集約データと集約データと名付けられた2種類のデータが提供された．

非集約データのためのデータ変換はとても単純である．問題1，2，4，および5では同じ非集約データが使用される．このデータはクリックストリームスターをジョインして得られたフラットなテーブルである．このテーブルでは各レコードはページ閲覧である．セッションにページ閲覧が複数個ある場合，セッションの属性は複数回繰り返される．同様に顧客情報はテーブルにおいて繰り返される．これら4個の出題問題で目標を定義するために，次の手順でテーブルに3個の2値属性を加えた．問題用にセッションを切り取る方法を示すために，表-1にセッション例を3個示す．

- 「問題1テスト集合」属性は，もしセッションが問題1と4のテスト集合にあれば，参加グループがこのページ閲覧を見られるかを表す．これはランダムに選択したセッションの途中にある切り取り点に基づいて定義されている．切り取り用に選択したセッションについて，我々は1から（セッション長−1）の間にランダムに切り取り点を生成した．長さ1のセッションについては，切り取り点は設定しなかった．
- 「問題2テスト集合」属性は，もしセッションが問題2と5のテスト集合にあれば，参加グループがこのページ閲覧を見られるかを表す．これはすべてのセッションで定義された切り取り点に基づいている．切り取り点は問題1と同様に生成された．
- 「セッション続行」属性は，問題1と4の目標として定義される．

問題3に対する非集約データ集合も注文行スターをジョインして生成されたフラットなテーブルである．各注文行はテーブル中のレコードである．注文ヘッダと顧客についての属性は複数回繰り返される場合がある．2値属性「1注文あたり平均12$以上支払う」を目標としてテーブルに追加した．この属性は顧客レベルで定義されている．

これら2個の非集約データ集合は生データを含んでおり，データマイニング用のデータ変換を行う能力がある人々にとって十分な情報を含んでいる．最初のデータ集合は注文情報を含んでおらず，2番目のデータ集合はクリックストリーム情報を含んでいないことに注意されたい．各参加グループは問題を解くのに必要と考えれば，それらのデータ集合をジョインすることが可能だった．

多くの研究者たち，特にデータマイニングアルゴリズムに取り組んでいる人たちが生データを変換（集約を含む）するソフトウェアを持っていないことを考え，我々はデータの集約版も提供した．集約データは，問題1と4用に1個，問題2と5用に1個，そして問題3用に1個と，3個のデータ集合から構成されている．これらのデータ集合は2個の非集約データ集合をマイニングに適切なレベルで集約することにより生成した．そのレベルとは，問題1, 2, 4, および5についてはセッションレベルであり，問題3については顧客レベルである．同時に，既存の属性を調べて新しい属性を追加した．たとえばセッションブラウザの種類，ブラウザの名前，および最もよく使われる3個のブラウザの種類を抽出した．問題1, 2, 4, および5に対する2個の集約データ集合において，各セッションは単一のレコードである．これら2個のデータ集合を生成する際，非集約データ集合の「該当するテスト集合に存在しない」と記されたすべてのページ閲覧は，集約オペレーションの前に取り除いた．問題3に対する集約データ集合において，各顧客は単一のレコードとして表される．

集約オペレーションにより問題1と4，および問題2と5において，それぞれ151, 153個の新しい属性が生成された．その例としては，データ集合の統計値に基づいて選択された各トップ製品についての閲覧数，種類に関する閲覧数，異なるテンプレートについての閲覧数，および最後のページに関する情報（ページ自体についての情報とその日時情報を含む）などがある．問題2と5に対しては，残りのセッションにおける各ブランド（ヘインズ，ダナキャラン，アメリカンエッセンシャルズ）についての閲覧数を示す3個の数値属性を定義した．これに加えて，残りのセッションにおいてどの3ブランドも閲覧されなければ真，そうでなければ偽となる2値属性も定義した．

問題3について集約データ集合を定義する際，クリックストリームは問題3を解くのに有用であり集約後のジョインが難しいと思ったのでクリックストリームデータと注文行をジョインした．このデータ集合に対する集約は最初にセッションレベル，次に顧客レベルと，2つのレベルで行われた．この結果，セッションあたり平均リクエスト回数，セッションの最初で参照元となるWebサイトページの上位5個，および日曜日に購買された製品のパーセントなど合計434個の新規属性が生成された．

	訓練集合		テスト集合	
問題	属性	レコード	属性	レコード
1: 非集約	217	777,480	215	164,364
2: 非集約	217	777,480	215	142,204
3: 非集約	232	3,465	-	-
1: 集約	296	234,954	296	50,558
2: 集約	299	234,954	296	62,913
3: 集約	518	1,781	-	-

表-2 データ集合の統計値

最終データスキーマと形式

データ集合は，データマイニングで広く用いられているC5形式（www.rulequest.com）を使ったフラットなファイル群として公開された．問題3は純粋な洞察問題であるため，訓練集合／テスト集合への分割は行わなかった．問題1と2には訓練集合とテスト集合があった．訓練集合には目標に関する情報があるが，テスト集合にはない．目標に関する情報が洩れることを恐れて，時間を十分使って訓練集合／テスト集合への分割を行った．ガゼルドットコム社から入手したデータは2000年1月30日から2000年4月30日まで（3カ月間）収集された．すべての問題に対し，訓練集合は2000年4月1日より前（2カ月間）のデータを使用した．問題1と2では同じ情報を使う場合があるため，それらのテスト集合は互いに排反であるべきである．我々は2000年4月14日より後（半月）のデータを問題1のテスト集合として使い，2000年4月1日から2000年4月14日まで（半月）のデータを問題2のテスト集合として使った．

表-2にデータ集合における属性数とレコード数を示す．表には問題4と5がないが，これは問題4が問題1と同じデータを使い，問題5が問題2と同じデータを使ったためである．問題2については，訓練集合には4個の目標属性があり，テスト集合にはこれらの値で定義される目標属性が1個だけある．

投稿の統計値

非開示の同意のもと，170回のデータアクセス要求があった．これらのうち，1個以上の問題に対して投稿した参加グループの数は31だった．各問題に対して受け付けた参加数を図-1に示す．コンテスト終了以来，研究や教育を目的として，データアクセスへの同意がクリッ

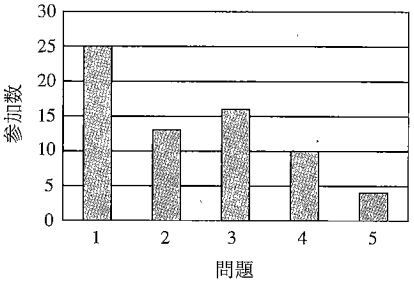

図-1 各問題に対する参加数

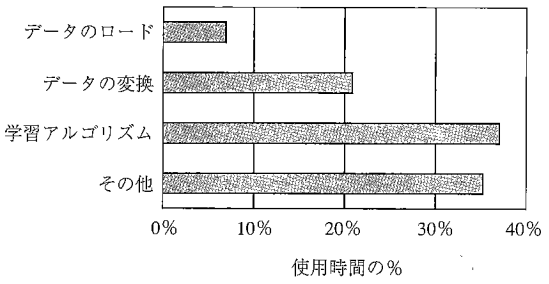

図-2 解析の各段階に費やした平均時間

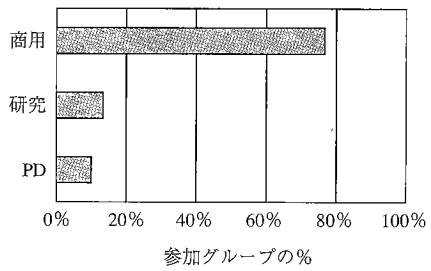

図-3 コンテストの各参加グループによって使用されたソフトウェアの種類．ただしPDはパブリックドメインを表す

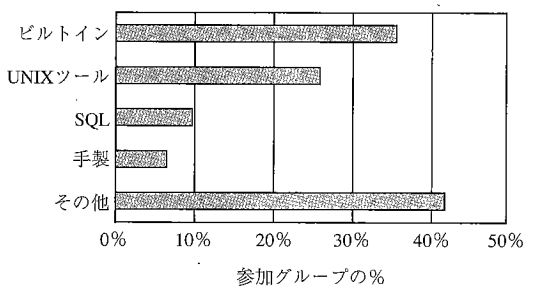

図-4 使用されたデータ処理ツール

クを通してさらに190回以上あった．

各参加グループには，今回のマイニング努力に関するアンケートへの回答を要求した．アンケートの質問は，用いた資源（例：分析の各段階で用いた人数と時間数），用いたソフトウェアとハードウェア，データ処理とデータ解析の両方において用いたデータマイニング技術と方法などを出した．この章で述べる統計値は，このアンケートへの回答に基づいている．

総計では，31の参加グループはデータの準備と解析に6,129人月を費やした．これは1参加グループあたり約200人月に相当する．投稿にあたって900人月を費やした参加グループも存在した．従事した人数は1人から13人の間だったが，ほとんどの参加グループは2人か3人で構成されていた．使用時間を細かく見た場合の平均値を図-2に示す．

多くの調査[4]に反し，データ準備とデータ変換に費やされた時間は30%に満たなかったことに注意されたい．考えられる理由は2つある．最初に，データはブルーマティーニ社のデータマイニング用に設計された統合eコマースシステムの中で収集されたため，比較的分析しやすい形式となっていた[5]．次に「使用データ」の章でも述べたように，我々は時間を十分使って，コンテストの参加グループが分析に使えるようにデータを変換し新しいフィーチャを生成した．

各参加グループが使用したデータマイニングソフトウェアの出所を細かく見たものを図-3に示す．注目すべき興味深い傾向として，KDDカップにおいて商用ソフトウェアの使用が増加していることが挙げられる．商用ソフトウェアあるいは所有権があるソフトウェアを使用するエントリーの割合は，44%（KDDカップ1997），52%（KDDカップ1998），77%（KDDカップ2000）と増えてきた．

コンテストの参加グループによって使用されたオペレーティングシステムの種類は，マイクロソフトWindows（54%）とUNIX（46%）が半々だった．UNIXを使った参加グループのうち，種々の商用UNIXが65%を占め，Linuxが残り35%を占めた．オペレーティングシステムではマイクロソフトWindowsとUNIXが半々だったにもかかわらず，ハードウェアは主にデスクトップPC（73%）であり，UNIXワークステーション（27%）は少数派だった．

データアクセスに関しては参加グループの32%がデータベースを使用し，68%がフラットなファイルを使用した．各参加グループによって使われたデータ処理ツールを細かく見たものを図-4に示す．この図よりほとんどの参加グループが分析ソフトウェアに付随するデータ処理ツールを使用し，このKDDカップのために新しいデータ処理ツールを開発する者は少なかった．

「使用データ」の章で述べたように，我々は集約データと非集約データを提供した．参加グループのうち，非集約データを使った者（41%）よりも集約データを使った者（59%）の方が多かった．このことより，多くのデー

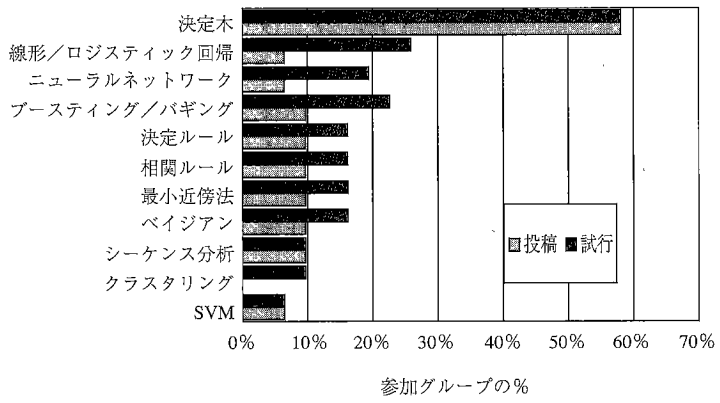

図-5 試行されたアルゴリズム対投稿されたアルゴリズム．ただしSVMはサポートベクタマシーンを表す

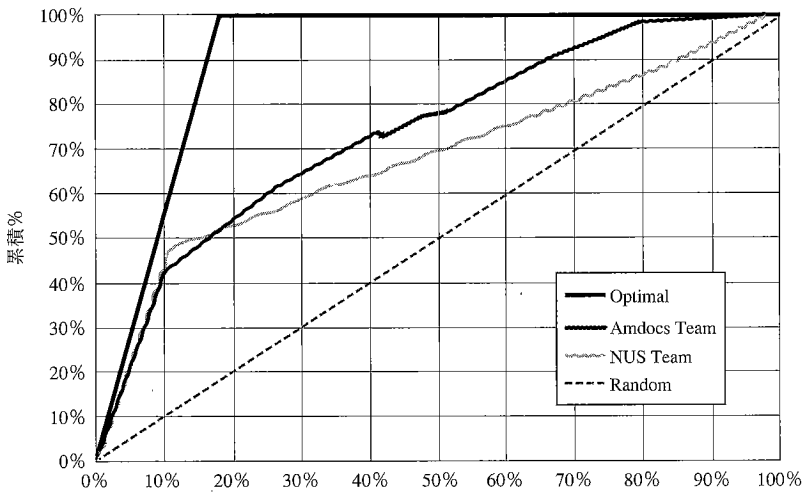

図-6 ページ閲覧数が5以上のセッションについての累積ゲインチャート

タマイニングツールはデータ集約に関して限られた機能だけしか提供していないことが推測される．

図-5に各参加グループが使用したアルゴリズム技術を示す．図にはアルゴリズムを試してみた参加グループの割合と，そのアルゴリズムを使って少なくとも1つの問題を解いた参加グループの割合が載せてある．図から分かるように，決定木は他を大きく引き離して最もよく使われた方法だった．50%以上の参加グループは決定木を使用して少なくとも1つの問題に対して解答を投稿した．

結果と洞察

本章では5つの問題のそれぞれについて結果を説明する．

各参加グループは実行可能な洞察を数多く発見した．それらはどの参照元Webサイトからきた訪問者が多額の出費をするか，どのページが閲覧中止の原因となるか，および人々のどの層が多額の出費をするかを含んでいた．背後に隠された原因を発見した途端，興味深いと思われる多くの洞察が自明となってしまった．そのような原因は通常時期かセッション長に関連していた．たとえば多くの参加グループは多額の出費をする人とガゼルドットコム社からの電子メールを希望する人との間に相関があることに気づいた．この回答を時間に関してプロットすると，劇的に変化することが容易に分かる．これはガゼルドットコム社がこの質問に関するデフォルトを2回変更したためである．多くのセッションは長さが1だったため，Webサイトを去る人を予測する問題は特に難しかった．このデータでは，検索ロボットがページを1つだけ見たセッションの割合は，全セッションの16%を占めていた．このことにもかかわらず，どの訪問者が実在する人間ではなく検索ロボットであるかを特定した参加グループは驚くほど少なかった．我々は短いセッションが除かれた場合の結果を調べ，各参加グループが各セッションに付与した予測確信順位が高い場合は正確な予想が可能であることに気づいた．

問題1（ページ閲覧の集合から訪問者が別のページを見るかあるいはWebサイトを去るかを予測）に関しては，正答率は77.06%から59.56%の範囲であり，平均値は73.14%だった．上位2者の差はたった0.10%であり，これは50セッションに相当する．実際，上位5者の正答率の差は統計的に有意ではない（信頼度95%は±0.37%に相当する）．この結果にもかかわらず，もし5ページ以上の閲覧を含むセッションを予想する問題に限定するならば，結果はより有意となる（1位と2位の差は1.5%であり，信頼度95%は±0.79%に相当する）．**図-6**より，上位2位までの参加グループのゲインチャートはこれらの長いセッションの10%までは最適ゲインにほぼ匹敵することが分かり，これは目標の43%を占める．最適ゲインはグラフの左端に位置する曲線で表されている．

問題4は問題1に該当する洞察問題だった．重要な洞察としては，検索ロボットとガゼル社のテストユーザは去り，Webサイト滞在の長さは参照元のWebサイトに関連する（マイクーポンからきたユーザはセッションが長

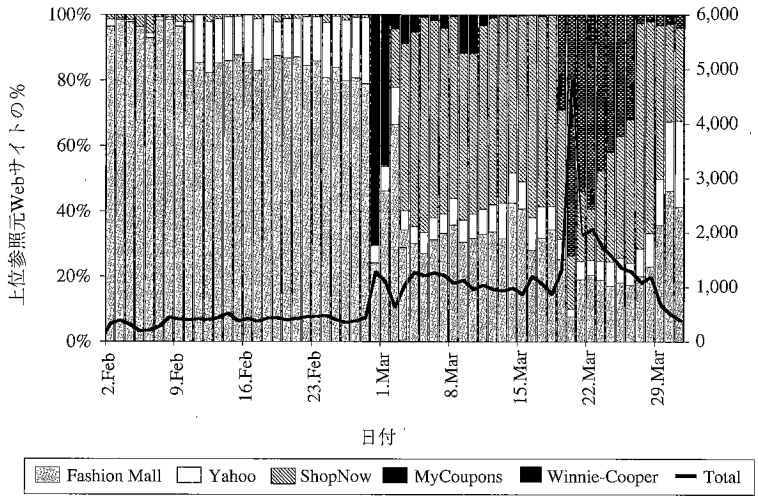

図-7 日ごとの上位参照元Webサイト

かったが，ショップナウからのユーザはすぐに去る傾向があった）などがあった．多くの参加グループは，2回目以降訪れたユーザが閲覧を続ける確率は初めて訪れたユーザの確率の倍であることに気づいた．特定の製品を見たために，ユーザがWebサイトを去ることがあった．これはWebサイトがそれらの製品を取り除くことを検討する可能性があるという意味で，実行可能な洞察の例となっている．他の実行可能な洞察の例としては，顧客の32%が消耗品を定期的に補充することを問うWebサイトのセクションを見た後で去ったことが挙げられる．多くの「発見」はWebサイトを去る確率がセッションで閲覧したページ数に応じて減少することに気づくことで説明できた．たとえば，「セッションで多くの異なる製品を見れば去る可能性は低い」という洞察はこの事実で説明できる．

問題2（ページ閲覧の集合から訪問者がセッションの残りで閲覧するブランドを予測）については，得点は60,956点から60,697点までの範囲であり，平均値は60,814.8だった．我々は問題1と同様に，上位の参加グループの差が統計的に有意でないことに気づいた．しかし問題1と同様に，評価をページ閲覧数5以上のセッションに限定した場合，リフト曲線が大変良くなった．最も良い予測因子の1つは参照元WebサイトのURLだった．ファッションモールとウィニークーパーからきた人はヘインズとダナキャランを好んで閲覧し，マイクーポン，トライポッド，そしてディールファインダからきた人はアメリカンエッセンシャルズを好んで見た．この結果をより詳しく調べると，アメリカンエッセンシャルズのブランドは主に靴下であり，これは安価であるため，しばしばクーポンの値段である10$より安くなることが分かった．ダナキャランが2月26日以降にならないと購入できなかったことに気づいた参加グループは大変少なかった．

問題3（Webサイトでの1回の注文において平均12$以上支払う訪問者の特徴付け）については，多くの興味深い洞察は単に時間に関連していた．たとえばTVコマーシャル，FRIEND販売促進，および事業の本格的立ち上げが起こった2月29日に，多くの活動が始まったことなどである．別の例としては参照元Webサイトのトラフィックが時間に応じて劇的に変化したことである（図-7参照）．この観察から得られた深い洞察は購入率に関連していた．Webサイトにおける訪問者全体の購入率はたったの0.8%に過ぎなかった．マイクーポンからの訪問者は8.2%の購入率だったが，支払い額は少なかった．一方，ファッションモールとショップナウからの訪問者は35,000人にものぼったが，購入率はたったの0.07%だった．多額の購買に関連する他の要素をいくつか次に示す．

- AOLの顧客でないこと（ガゼルドットコム社のWebサイトは，AOLのブラウザウィンドウではきれいに表示されない）．
- 印刷された広告を見てきた訪問者である．
- 収入が大変高いか大変低い．
- 米国北東部に住んでいる．

得られた教訓

KDDカップは，データマイニングの研究者とデータマイニングのユーザの両方に対して参加グループと開催者が直面する問題を明らかにするという意味で素晴らしい手段である．ここでは得られた主要な教訓を述べる．

最も重要な教訓は，興味深さの評価指標が正答率や得点（問題1, 2）だけだとしても，人間がKDDプロセスの重要な一部であることだ．完全に自動化されたツールは，優勝グループが多段階モデルを生成し（KDDカップ2000優勝者レポート[2),8)]参照のこと），検索ロボットや企業のテストユーザを特定し，そして追加フィーチャを構築した原因である人間の洞察に決してかなわない．人間の理解が重要であることは，試行されたアルゴリズムと実際に使用されたアルゴリズムを見ても明らかである．決定木は最も頻繁に試行され投稿においても最も頻繁に用いられたが，ニューラルネットワーク，ロジスティック回帰，およびクラスタリングは使用／試行率が最も悪かった．興味深い結果を発見したと思った多くの参

加グループは，本質をつかんで背後に隠れている真の原因を発見するために時間を割かなかった．洞察問題においては，多くの初期相関が自明であり産業界のユーザにとって興味深くない（例：多額購買の閾値である12$以上の商品を購入する者は確かに多額購買者である）ため，プロセスを反復することはさらに重要だった．興味深いと思われた多くの洞察は支持度がきわめて低かった．たとえば，いくつかの参加グループがショップナウからきた購買者はすべて多額購買者であると述べた．このことは事実であるが，支持度はたったの6人だった！　人間が関与する必要があるため，有用な洞察を発見するにはたとえば何百時間というように時間がかかる．

Webサイトの変更により，興味深い影響とバイアスが生じた．特別なマーケティングイベントに関する文書を無視したグループは成績が悪かった．時間はきわめて重要な属性であり，Webサイトと製品への変更は考慮する必要がある．1つの立場として，ある参加グループは問題が現実的すぎるので我々に簡略化すべきだと主張した．我々の問題は難しかったが，実世界の設定に基づく実問題である．結果は，特に長いセッションに対して有意なリフトを示し，多くの洞察がきわめて興味深いうえに実行可能だった．1クリックセッションの多くについては精度良く予測することが不可能だったが，予測確信順位が高い場合，特に長いセッションに対しては，予測結果は大変よかった．

データはブルーマティーニ社のアプリケーションサーバを通して収集され，標準的なWebログは使わなくて済み，購買をクリックストリームに関連付けることが可能だった．データ収集システムはURLに加えて表示した製品に関する情報を保存するため，情報はWebサイトの変更に対して比較的安定していた．このようなデータは解析にあたって比較的少ない労力で多くの情報をもたらす．さらにアクシオムの属性を加えたことにより，洞察を導くのが確かに楽になったはずだ．このようにWebログよりも有利だったにもかかわらず，検索ロボットとテストユーザの特定はやはり難しかった．

将来のKDDカップの開催者に，いくつか提案したい．KDDカップ開催のボランティアを引き受ける前に，多大な労力が必要であることを覚悟すべきだ．推定では我々はデータを入手し，クレンジングし，隠匿し，変換し，Webページを設定し，法的合意文書を作成し，そして結果を評価するのに合計800時間を費やした．十分な時間を使ってデータ隠匿とデータに隠されたリークを特定すべきだ（収集プロセスにおいて，結果として現れるため目的属性を明らかにしてしまう「暴露属性」を放棄する）．たとえば，我々のシステムはセッション長を保存したが，我々はデータを切り取った後でこの属性を再計算する必要があった．さもなければセッション長はWebサイトを去るかという目的属性を明らかにしてしまうからだ．前年までの失敗を見てきたため，今回我々は大変注意深くリークを除いた．それでもランダム化とクッキーの隠匿に関して間違いがあったため，初期段階において2度データを再公表しなければならなかった．我々は多くの時間を費やして背景知識を与え属性を説明するなどの導入説明を書いた．それでもFAQを作成する必要があり，質問数はコンテストの終わりに67となった．参加グループにはデータ公開直後と投稿前と，質問時期を2度与えた．参加グループに解析を始めさせる点と我々が時間をよりよく配分する点の両方において，この方式は有用だったと思っている．評価は時間がとてもかかり，特に重み付けされた洞察のリストの作成と洞察の妥当性検証が大変だった．参加グループには産業界のユーザ用のレポートを書くように要請したが，レポートを読んだ後で我々は多くの著者たちが産業界のユーザと話しすらしたことがないと感じた．良かった点としては，我々は知らなかったデータについて多くのことを学び，結果を示したり結果をより多くの人たちに分かりやすくする素晴らしい方法を見たことが挙げられる．

謝辞　データを提供していただいたガゼルドットコム社に感謝する．アクシオム社は公開情報を無償で提供し配布することを認めていただいた．ブルーマティーニ社の小売り専門家であるCatharine HardingとVahe Katrosには，投稿の査読を手伝っていただき発見されたパターンを説明していただいた．パーデュ大学のSean MacArthurには得点計算プログラムのコーディングを手伝っていただいた．

参考文献

1) Kohavi, R. and Provost, F.: Applications of Data Mining to E-commerce (editorial), Special Issue of the Int. J. on Data Mining and Knowledge Discovery, http://xxx.lanl.gov/abs/cs.LG/0010006 (2001).
2) ACM SIGKDD Explorations Homepage, the Information for Authors Link: http://www.acm.org/sigkdd/explorations/
3) Brodley, C. and Kohavi, R.: KDD-Cup 2000 homepage, http://www.ecn.purdue.edu/KDDCUP/ (2000).
4) Piatetsky-Shapiro, G. et al.: An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications, Proc. KDD-96, pp.89-95 (1996).
5) Ansari, S., Kohavi, R., Mason, L. and Zheng, Z.: Integrating E-Commerce and Data Mining: Architecture and Challenges, Proc. WEBKDD'2000 Workshop, http://robotics.stanford.edu/users/ronnyk/WEBKDD2000/ (2000).
6) Kimball, R., Reeves, L., Ross, M. and Thornthwaite, W.: The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses, John Wiley & Sons (1998).
7) Kimball, R. and Merz, R.: The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse, John Wiley & Sons (2000).
8) Kohavi, R. et al.: KDD-Cup 2000 Organizers Report: Peeling the Onion, ACM SIGKDD Explorations, 2 (1) (2001).

(平成13年3月31日受付)