情報システムと社会環境　83－3
（２００３．　３．　１４）

アカウンティング情報を用いた利用者の専門性分析

神奈川大学理学部情報科学科

田代 淳一[†]　　堀 幸雄[‡]　　後藤 英一[§]

現在アカウンティング情報は主に利用者への課金目的に利用されている．しかしアカウンティング情報からはその使用傾向や行動パターンが分析可能であり，個人の知識管理に利用できると考えられる．本研究ではアカウンティング情報を用いて利用者の動向分析を行い利用者の類似性と専門性の評価を行なった．クラスター分析を用いて類似性の判定を行なった結果，同じ専門性を持つ利用者を正しく分析し，その要素を限定できることを確認した．

Expertise analysis of the user using accounting information

Department of Information Science,
Faculty of Science, Kanagawa University

Junichi Tashiro[†]　　Yukio Hori[‡]　　Eiichi Goto[§]

Accounting information is mainly used for a user's fee collection purpose. However, it is thought that the use tendency and action pattern can be analyzed and it can use for knowledge management of an individual from accounting information. In this study, trend analysis of a user was performed using accounting information, and evaluation of a user's similarity and speciality nature was performed.As a result of judging similarity using cluster analysis, the user with the same speciality nature was analyzed correctly, and it checked that the element could be limited.

1　はじめに

本研究着手にあたる背景は，次の２つの点に着眼したことである．1. アカウンティング情報が利用者に与える有用性が極めて少ない.2. IT 技術者の計算機スキルを体系化した指標がいまだ統一されていない．

アカウンティング情報は容易に取得することができ，個人の計算機に関する情報が網羅されているにも関わらず利用者向けの情報ではなく管理者向けの情報として認識，利用されているため，利用者に対して特に有用されてないのである．

また計算機スキルの指標の統一されていないことも大きな問題点であり，統一化は非常に重要なことである．近年 IT 産業の売上げ及び利益がハードからソフト・サービスにシフトするに伴い，戦略的・体系的なスキル管理・育成の重要性が高まってきている．メインフレームの時代には，スキル伝承手段として OJT(On the Job Training) が機能していた．しかし各個人のスキルの明確化が情報サービスを提供している企業にとって課題となっている今日においては，従来のスキル管理を見直す必要がある．すなわち，市場において求められるスキルが多様化・深化するに伴い誰が何をできるのか客観的に指標化し各個人もまた把握することがスキルアップのうえで重要となってくるのである．国の対応としても 2002 年 12 月に経済産業省が IT サービスに必要とされる実務能力を体系化した指標である「IT スキル標準」[1] が策定段階であり，国内における能力の指標化の整備が必須課題であることが明らかである．

また商業においても購買履歴といった顧客情報をマーケティングや販売戦略に活かすようにアカウンティング情報も利用者情報としての価値を有用化し組織の活性化につなげたいと著者は考える．

本研究の目的は，上記のような側面を統合的に考慮しながらアカウンティング情報から利用者の専門性が分析できることを示すことである．

2　アカウンティング情報と専門性

2.1　アカウンティング情報とは

アカウンティングシステムは，主にユーザに課金するために，システムリソースの使用量を記録するよう

[†]tashijun@goto.info.kanagawa-u.ac.jp
[‡]horiyuki@goto.info.kanagawa-u.ac.jp
[§]goto@goto.info.kanagawa-u.ac.jp

に設計されている．一般的にUNIXでは，ユーザベースのプロセスの情報を記録・管理している．すなわち，オペレーティングシステムでは，実行される各プロセスについて，ユーザ含む統計値を含むことにより，システムの使用量が記録される．アカウンティングシステムによって収集されたデータは，システムパフォーマンスの監視にも使用することができる．アカウンティングレコードには，システムで実行する各プロセスに関する次のようなデータが収められている[2]．

- コマンドイメージ名
- 使用したCPUタイム
- プロセスが完了するまでの経過時間
- プロセスが開始された時刻
- 対応するユーザIDとグループID
- 存続時間でのメモリ使用量
- 読み書きした文字数
- 読み書きしたディスクI/Oブロック数
- プロセスが起動されたTTY
- プロセスに関連するアカウンティングフラグ
- プロセスの終了ステータス

上記のデータを用いて，専門性分析を行う際にいかなるデータ処理が必要かについて次節で論ずる．

2.2 専門性分析手法

前節で述べたデータをアカウンティング情報から取得し，とくに使用したプロセスを重み付けを行う．このとき成分をv，サンプルサイズをn，重み付け係数をwとし，

$$v_j = \sum_{i=1}^{n} w^{(i-1)}$$

で表した．wは，0.99とした．これはプロセスの単純な線形総和の場合，特徴空間で出現頻度の莫大な差があるためこの範囲を0～100までの間に正規化している．すなわち，このデータがアカウンティング情報からクラスター分析を行う際の対象である．クラスター分析では，クラスタ間の距離(非類似度)関数に基づき，最も距離の近い二つのクラスタを逐次的に併合する．そして，この併合を，すべての対象が一つのクラスタに併合されるまで繰り返すことで階層構造を獲得する．この階層構造はデンドグラムによって表示する．デンドグラムは、各終端ノードが各対象を表し，併合されてできたクラスタを非終端ノードで表した2分木である．非終端ノードの横軸は，併合されたときのクラスタ間の距離を表す．クラスタを形成する際の対象間の距離を求めるにはいくつかの手法があるがウォード法(Ward'smethod)[6]を適用した．ウォード法は以下のような一般式で表せる．

$$D(C_1, C_2) = E(C_1 \cup C_2) - E(C_1) - E(C_2)$$

$$\text{ただし，} E(C_1) = \sum_{x \in C_1} (D(x, c_i))^2$$

ウォード法は，各対象から，その対象を含むクラスタのセントロイドまでの距離の二乗の総和を最小化しクラスタを形成していく手法である．これはクラスター分析の中でも階層的方法の一つであるが，実用性が高い手法であり，鎖効果が起きにくい．鎖効果とは，ある一つのクラスタに対象が順に一つずつ吸収されてクラスタが形成されていく現象である．このようなデンドグラムが得られた場合には，どの距離で切ってもあるクラスタとその他の対象一つずつで構成されたクラスタに分かれることになり，グループに分けたことにならない．

ウォード法により形成されたクラスタは，クラスタ間の結合距離の近いもの同士で成り立っているが本研究では要素の次元数が1000以上と膨大なためクラスタ間のプロセス成分の相関関係を求めるのは難しい．そこで，クラスタ間に主成分分析を行い軸の回転を利用して対象を記述するのに最小次元を獲得，すなわちクラスタ間で特化しているのか成分を抽出する．

専門性の定義は，クラスター分析によって分類されたクラスタで代表となる成分やクラスタ間で相関の強いものとする．代表となる成分とは，クラスタ内のユーザが共通して使用したプロセスの中でも特に頻度の高いものことであり，相関が強いものとはユーザの第一，第二主成分得点である傾向を示している成分のことである．

3 アカウンティング分析実験

3.1 方法と目的

前節で述べた手法を用いて，アカウンティング情報から利用者の類似性を見つけ，専門性となりうる要素を限定することを目的とする．

今回は，神奈川大学の教育用共同計算機システムで，平成14年3月から平成14年12月までの本校の利用者(理学部情報科，化学科，生物科，経営学部国際経営学科，学生と教職員)のアカウンティング情報を対

象とした. クラスタ分析はウォード法の自作実験プログラムを用いて, ユーザ ID 同士のクラスタリングを行う. 最終的に一つのクラスタに全てのクラスタが統合されたら, 終了する. その結果を用いてデンドグラムを形成する. また, その後適当な部分クラスタを選択し, そのクラスタ内での成分値を実験プログラムで主成分分析を行い, クラスタにおける成分の主成分得点係数と対象ユーザの主成分得点を出し, 変量とユーザについて散布図を作る.

3.2 実験結果および考察

上記手続きを実行した結果, 全てのケースにおいて, デンドグラム, 散布図を得ることができた. その中で特に注目すべき結果を図 1〜図 5 に載せた. これは情報科学科 3 回生のクラスター分析と主成分分析によるデンドグラムと散布図である. 特に 3 回生を利用した理由として, 授業カリキュラムで主に情報系専攻科目が大半を占めており, プログラミングを用いた演習や実験等が多く, 他学年の情報科学科あるいは他学科に比べ計算機システムの利用頻度も高いからである.

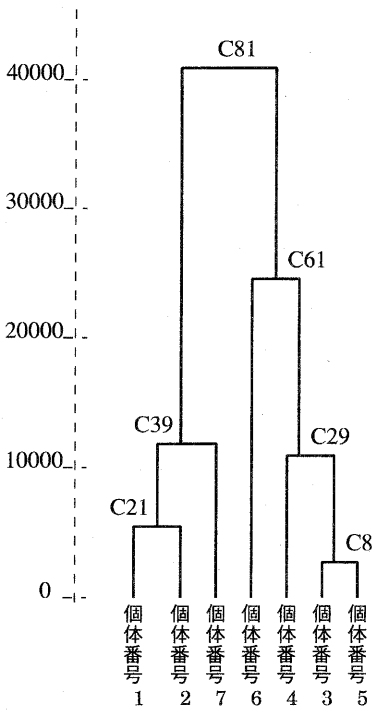

図 1: クラスタ (C81) のデンドグラム

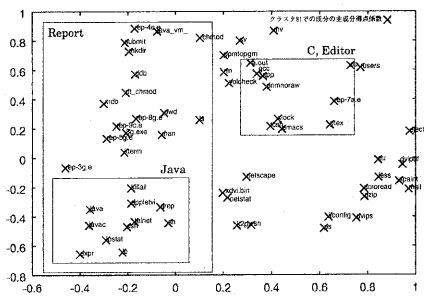

図 2: クラスタ (C81) の変量プロット

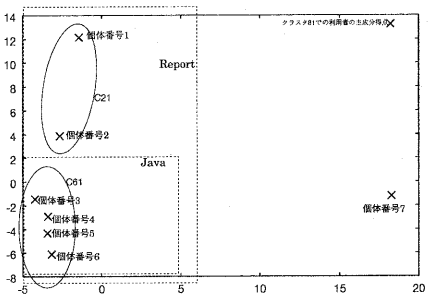

図 3: クラスタ (C81) のサンプルプロット

図 1 のデンドグラムにある個体番号 1〜7 は, 次のサンプルプロットと変量プロットの個体番号のことである. ここで変量プロットについては, 主成分分析対象である成分からあらかじめ説明変量となりうる成分を著者の勘案で選別し, どの利用者でもログインしているときに用いるであろう成分については除いてあることを述べておく. 変量プロットは第 1 主成分得点係数を横軸, 第 2 主成分得点係数を縦軸とした成分の散布図である. またサンプルプロットは, 各変量に割り当てられた主成分得点係数を用いて利用者の主成分得点を出したものであり, 横軸を第 1 主成分得点, 縦軸を第 2 主成分得点とした散布図である.

この変量プロットとサンプルプロットから興味深い発見が幾つかある. まず変量プロットを見ると, このクラスタにおいてある程度成分がその専門性毎に分かれている. 横軸の第 1 主成分得点係数からの側面で見ると正の領域には C 言語関連の cpp,a.out, それに gcc といった成分がある. 負の領域には逆に Java 言語を用いる際のプロセスである java,javac それに java アプレットを実行する際の appletvi がある. 他の成分も含まれているが, 明らかに C 言語系の成分と Java 言語関連の成分とで正負に分かれているところから, 第 1 主成分

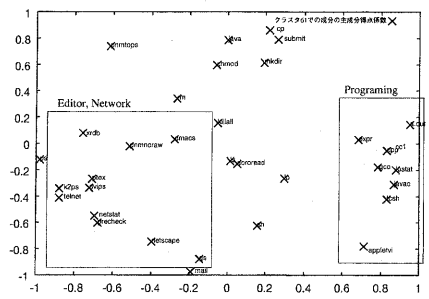

図 4: クラスタ (C61) の変量プロット

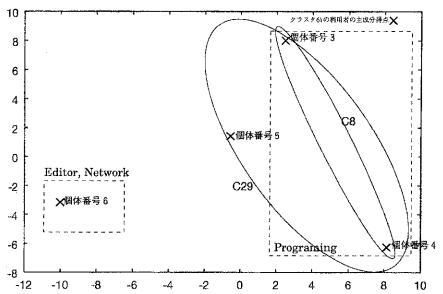

図 5: クラスタ (C61) のサンプルプロット

得点係数は Java 言語と C 言語を大別する値であると解釈できる, すなわち利用者が Java 言語と C 言語のどちらの利用傾向があるかを識別しているのである. また, 第 2 主成分得点係数では, 正の領域に先頭が rep で始まる成分が顕著に分布している. これは, 計算機システムを用いた演習課題を課する講義の提出用実行プロセスである. 特にこの講義では通常, プログラングは C 言語を用いる学生が既知であり,C 言語関連との正の相関があることは多いに考えれられる. 従って横軸を基本として見ると第 2 主成分得点係数は, 提出用実行プロセスに用いた言語が Java 言語か C 言語かを判別する成分と考えることができる. また Java 言語関連の成分,C 言語関連の成分は第 2 主成分得点係数からみても第 1 主成分得点係数からみたときと同じ関係にあり, 利用者の Java 言語か C 言語の専門性を大別するのは第 2 主成分得点係数からも全く言えることである. すなわちこれらを統合的に考慮すると, 第 1 主成分得点係数と第 2 主成分得点係数が正の場合, 普段比較的 C 言語を用いて演習などを行っていると考えることができ, 第 1 主成分得点係数, 第 2 主成分得点係数ともに負の場合, 日常的に C 言語よりは Java 言語のほうが利用頻度が高いことが言える.

上記の変量プロットの結果をサンプルプロットに反映させた結果, 図 2 のようになる. この結果を見てわかるようにクラスタ同士で散布がわかれており他のクラスタに比べ比較的クラスタ間のプロセスの専門がでている傾向があると考えられる. クラスタ分析によりクラスタ化されたものが類似性が近く, 得る専門性がクラスタによってちがうという結果のよい例といえる. 個体番号 1 と 2 はクラスタ 21 で, 個体番号 3〜6 はクラスタ 61 で結合されていて, クラスタ 61 は, どの利用者も第 1, 第 2 主成分得点がともに負の領域にある. これは先の変量プロットをみてもわかるように,Java を普段から利用していることが高いと思われる. したがってこのクラスタ 81 内においてはクラスタ 61 は Java の専門性があると断定できる. また, それより第 2 主成分得点が正の領域では, クラスタ 21 の利用者が分布されている. これはクラスタ 61 と比べ, 横軸第 1 主成分得点に関してはほぼ同じ値を取るものの, 第 2 主成分得点に関しては, 正負で全く反対の値をとっている. すなわち, クラスタ 21 は Java を普段から他の成分と比較して多く使っているという専門性はいえないが, レポート用の実行プロセスに関してはクラスタ 61 に比べ多く行っていることが明らかであり, 授業の課題に対する積極性が高い傾向がわかる. 今回は, レポート課題の実行プロセスだったため, 専門性とは言いがたいが課題に対する積極性が高いということも一つのアビリティとして考えれば, 一専門性として捉えることもできる.

ここで, 下位クラスタの比較をしてみる. 図 4, 図 5 はクラスタ 81 の下位クラスタのそれぞれ変量プロットとサンプルプロットである. クラスタ 61 はクラスタ 81 で, 特に Java の専門性が高かった. クラスタ 61 での変量プロットでは第 1 主成分は, 正の領域に主に gcc,a.out,cpp や javac,appletvi といった C 言語や Java 言語を用いる際のプロセスが分布しており, 負の領域には dvips,emacs や ptex といった文書作成の Latex に用いるプロセスやエディタのプロセスが分布している. また,netscape や netstat といったプロセスも負の領域にあり, 文書作成などのエディタとの相関があると考えられる. この変量の主成分得点係数を用いたサンプルプロット図 5 を見るとクラスタ 61 もクラスタ

	第一主成分 (%)	第二主成分 (%)
クラスタ 81	35.17	20.56
クラスタ 61	44.11	29.40

表 1: 主成分の寄与率

81 のサンプルプロットと同様, 構成しているクラスタ毎に散布が分かれている. 個体番号は図 1 のデンドグラムと図 3 のデンドグラムとも対応している. すなわちクラスタ 61 の結合でみると個体番号 3 と個体番号 4 は第 1 主成分得点が特に正の値を示しており, 図 4 の変量プロットから考えると主にプログラミング言語関連のプロセスが高い頻度で利用されてることが推測でき, 特に他のプロセスに比べプログラミングの専門性が高いと思われる. また個体番号 6 は明らかに負の領域に位置しており, 他のプロセスに比べエディタやネットワークのプロセスが高く, 特に頻度が高いことが分かる. また, 全ての主成分分析対象成分の中での寄与している割合をみる寄与率というのがある. これをクラスタ 81 とクラスタ 61 の主成分得点係数に対応させて第 2 主成分までを表 1 に示す. これを見ると, クラスタ 81 は, 第 1 主成分が 35%クラスタ 61 では第 1 主成分が 44%となっており, クラスタ 61 のほうが第 1 主成分が全体から見た影響度が高い. しかし寄与率を累積した場合, 両方のクラスタはともに第 2 主成分までが 50%を越えているものの, 元データの全体を集約しているといえる 80%を越えていないため, 全ての情報を集約はしていないことがわかる. この点から累積寄与率が 80%を越えている主成分までを含めた測定が次回以降, 必要となる.

アカウンティング情報を知識として捉え, 利用者の専門性かつ類似性の判定ができるかを検証した. クラスタ分析と主成分分析を融合させることでトレンド発見できることは一般的なことであり本研究でもこの手続きを適用したわけだが, 必ずしもどのクラスタにおいても共通の傾向が発見できたとは言いがたい. しかし, 特定の部分クラスタによっては第二主成分までの散布を見ると似た傾向が出ているクラスタもあることから, 距離とコマンドの使用傾向は相関性が高い.

4 今後の課題

本研究は, アカウンティング情報から計算機の利用履歴を分析し個人の計算機スキルを導き出し, 計算機スキルの体系化を行った. すなわちアカウンティング情報はアカウントを持った利用者の知識の潜在的に持っていたのである. この知識の応用としてナレッジコミュニティへのアプローチを課題として考えてみたい. アカウンティング情報から専門性が導き出せることは, ナレッジコミュニティにおいて専門家の特定する際の要素として用いることができないかということである. 以下にそのナレッジコミュニティについて概略を説明する.

4.1 ナレッジコミュニティ

デフレの進行や B2C 市場を中心とした IT バブルの崩壊, 中国メーカーなどが低価格攻勢を強めるなか,IT 技術を業務効率化, 合理化に用いることはもはや必然的なことであり他社との差別化にはならない. 今や企業競争力の底上げには社内外の知識財産を一元管理し重要な経営資源として経営に活かすことが今企業がすべき重要課題の一つである. すなわち財務諸表の資本と並んで知識資本の良し悪しが企業の競争力の決め手となる時代が始まったのである [7]. モノ作り大国であったわが国企業のホワイトカラーは, 米国企業に比べて生産性が三分の一程度であるといわれていた. しかし知識社会が到来するに及んで, 生産性の低いわが国の多くのホワイトカラーもいかに効率的に知識資産を経営資源に活用するかが問われているのである.

このような背景から最近, 企業でデジタルネットワークを活用して社内外の専門家に技術相談できる仕組みが注目され始めており, 導入企業も相次いでいる. このような仕組みを一般的にナレッジコミュニティという. 具体的には, 社内エキスパートや参加者の知識, 得意分野をデータベースに登録し, 質問を電子掲示板に書き込むと社内専門家や技術者がメールで回答するという流れである. このような仕組みを企業経営に活かす経営戦略手法をナレッジマネジメントといい企業の知識資産を活用においてのトレンドとなっている. ナレッジコミュニティは社内に埋もれている個人の潜在的な知識 , ノウハウを共有可能となり会社の資産として活かすことができる. また同時にこれをイントラネットで構築することで社内で誰がどんな知識を持っているのかということも一元管理できる. 特に率先してナレッジコミュニティを導入している先進企業においては, この知識管理の仕組みに人事考課の評価システムなどを併せて運用している. すなわち, どれだけナレッジを企業に還元したかということを人事考課制度として評価する仕組みである. このコラボレーションにより, 社員がナレッジを表出するモチベーションにつなげているのである.

4.2 おわりに

知識社会における本研究の位置付けは, 教育用計算機ルームでの計算機スキルの体系化と専門家検索である. 例えば, 大学の情報系の学科で学内教育用計算機システムを用いるときに利用者間の使用傾向の類似度を見ることができる. またある特定のコマンド郡 (言語関連, システム, ユーティリティ) などについて特に専門性が高い利用者を検索することも可能であり, ティー

チングアシスト (TA) 選出や, 技術的な問題が発生したときの要員検索といった人員配置, 利用者が自分の能力を客観的に知ることに有効である. こういったアカウンティング 情報が知識社会に与える有用性を本研究では, 特に取り上げたのである. これを具体化する提案として, 本研究をシステム化して大学の計算機への導入がある. 導入された環境下で利用者はコマンドラインから実行し , その時点での自分にはどのような専門性があるか, また他の利用者がどういった専門性があるかを確認でき, その分野に関する問題, 疑問がある場合に誰がその知識があるかを検索可能となり現在の Q&A 方式のナレッジコミュニティとは別の角度からの組織内の専門家に技術相談できる仕組みが構築できるであろう.

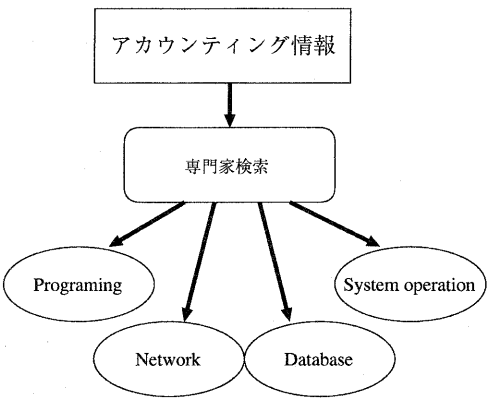

図 6: アカウンティング情報からの専門性発見

参考文献

[1] 経済産業省 IT スキル標準 ver1.0(2002)

[2] AEleen Frisch 著. 谷川哲司 監訳. 黒岩真吾 株式会社ユニテック 共訳: UNIX システム管理 改訂版オライリー・ジャパン (1998)

[3] Everitt,B.S.:Cluster Analysis,Edward Arnold, third edition (1993)

[4] Jain,A.K.and Dubes,R.C.:Algorithms for Clustering Data,Prentice Hall (1988)

[5] Jain,A.K.,Murty,M.N.and Flynn,P.J.: Data-Clustering:A Review,ACM Computing Surveys,Vol.31,No.3 (1999)

[6] 神嶌敏弘: データマイニング分野のクラスタリング手法 (1). 人工知能学会誌, Vol.18 No.1 pp.59-65(2003)

[7] 山崎秀夫 著: 未来型組織を支える企業ナレッジポータル. 野村総合研究所 (2002)