

発生頻度に頼らないユーザー行動データの自動処理 Automated processing of user activity data without relying on the frequencies

太田 武昭*¹
Takeaki OTA

清水 英則*¹
Hidenori SHIMIZU

*¹ 日本ヒューレット・パッカード株式会社
Hewlett-Packard Japan, Ltd.

We propose a scheme for timely processing user activity data using the event processing model. Our processing flow firstly filters out user's ordinary activities using traditional data-mining or heuristic approaches. Then, for other rather rare and minor but potentially very important cases for the user, the flow tries to provide appropriate qualitative interpretations of user's activities, by looking up cross-user heuristic based reasoning rules. Assuming that users devote less attention to the familiar activities for which they are skilled, we can put more service and/or human-support resources to those less usual yet more qualitatively important situations. The actual work and benefits of the scheme are explained using the real user access data of the enterprise intranet-SNS web site.

1. はじめに

情報通信技術の発展とともに、今日では各種の情報通信サービスを時間や場所を問わずに様々な環境から利用する利用形態が増加してきている。通常、サービス利用者側は日常生活を送りながら、適切な時間に適切な頻度で特定または複数のサービスを利用することとなる。それ故、サービス提供側が利用者個人のおかれた状況を適切に見定め、適切な内容と品質のサービスを提供していくことはますます重要になると考えられる。

特定のサービス利用者がサービスにアクセスしにくる状況を推測する際、当該利用者または類似者の過去の行動履歴と現状を比較することが有効であろうと考えられる。そこで今回、企業向け情報共有システム [宮内 09]で報告されているウェブサイトのアクセスログデータを分析調査したところ、発生頻度が高い行動以外で、発生頻度が少ない事例でも利用者にとって重要な状況がある場合が判明した。本稿では、これらの実情も考慮して検討した、利用者のアクセス状況推測の一手法を紹介する。

2. ユーザー行動データの自動処理の概要

2.1 処理データの形式および処理の目的

提案方式では、ある特定の利用者($User_i$)の行動に伴い発生したイベント($Event_m$)の重要度($Importance$)を可能であればその理由($Reason_x$)とともに提示することを目的とする。ここでイベント($Event_m$)は、必須項目であるイベント発生時間($EventTime$)、イベント発生源識別子 ($EventSourceId$)、利用者識別子($UserId$)、行動内容($ActivityType$)、ならびに任意項目である行動詳細($ActivityDetails$)より構成される。

上記の項目は十分に基本的な内容であり、ユーザー行動に関連する様々な種類のデータに汎用的に適用可能であると考えられるが、ウェブアクセスログへの適用について後述する。

2.2 処理方式

図1に提案手法が用いるイベントデータの処理方式を示す。

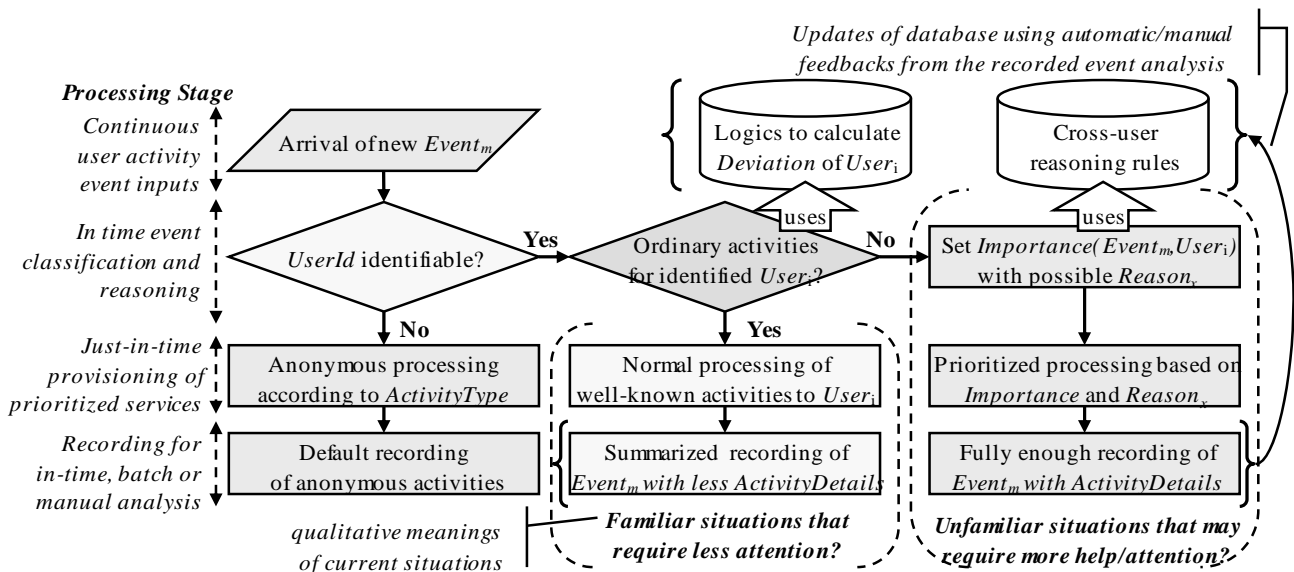


図1 ユーザー行動データに関連するイベントデータの重要度判定処理方式

処理の概要であるが、まず複数のユーザーの行動に関連するイベントデータが、イベント発生に応じて継続的に入力される。

次に入力が特定のユーザーに関するイベントであるかどうかを判定する。ユーザーが特定できない場合、匿名ユーザーへの規定の優先度でイベントに関連するサービスの提供処理を行い、入力イベントおよび関連処理の内容をログに記録する(図中左)。

ユーザーが判定可能な場合には、当該ユーザーが日常的に行っている行動を示すイベントであるかどうかをユーザーの過去の行動履歴データベース等を用いて判定する。ユーザーが日常的に行っている行動である場合には、該当ユーザーに対する通常の優先度でイベントに関連するサービスの提供処理を行い、入力イベントおよび関連処理の内容は要約された形式でログに記録する(図中央)。

一方、過去に当該ユーザーが行った頻度が少ない、あるいは初めての行動に関するイベントであると判明した場合においては、事前に複数ユーザーを対象にして構築しておいた理由付けデータベースに蓄えられているヒューリスティック等を用い、入力イベントに対するユーザーの行動理由・状況の推測を試みる。更に、判定された重要度と理由に応じた形でユーザーに提供するサービスの処理優先度を高める。また、入力イベントおよび関連処理の内容は詳細な形式のまま記録する(図中右)。

なお、ここでの重要度の計算では、 $Importance(Event_m, User_i) \approx Attention(Event_m, User_i) \times Importance(Event_m) \approx Deviation(Event_m, PastActivities(User_i)) \times Importance(Event_m)$ 。という近似が有益だと仮定の下、特定ユーザーの過去の行動との相違とイベント自体の重要度の積で代用している旨了承されたい。

他方、サービス提供ログに記録された各種情報については、後にバッチ形式や人手で処理し、履歴データベースや理由付けデータベースのフィードバック更新等に用いられる。この際、特に発生頻度の少ないデータについては詳細な関連情報を記録しておくことで、より適切なヒューリスティックルールの作成を助けられることが望める。一方、各種センサーやマルチメディア形式などのデータ量の大きな情報の記録においては、一般的な利用状況を要約して記録することで管理コストが低減できる。

3. 実データを用いた適用可能性の検討

提案手法の具体的に適用事例として、ユーザー行動データに各種サービス提供サイトにおけるウェブアクセスログの情報を利用し、ユーザー状況をアクセス時に判定する方式を検討する。

3.1 分析対象としたウェブアクセスログの概要

[宮内 09]で報告されている企業向け情報共有システムにおいて、掲示板/Blog/WiKi/ファイル共有等の各種サービス提供に用いられている社内イントラネット向けウェブサイトのアクセスログデータを分析した事例を報告する。

ここで対象とするデータは一般的なウェブサーバーのアクセスログであるが、社内認証システムと連動した非匿名での利用が義務付けられており、数千人規模のユーザーの個人毎の利用状況の分析が可能である。また、ログに記録されているブラウザのクライアント IP アドレス情報を用いることで、大まかなアクセス位置情報(社内のオフィスのフロアレベルや社外からのアクセスなど)を入手できる。

3.2 利用者データの実例

今回は 2007 年度一年分の過去の利用者のデータを用いて、利用頻度の絶対数が異なる利用者を複数名選択し、インタビューを含めた分析調査を行って、日常性の判定方式及び理由付けデータベース向けのヒューリスティック規則の検討を行った。

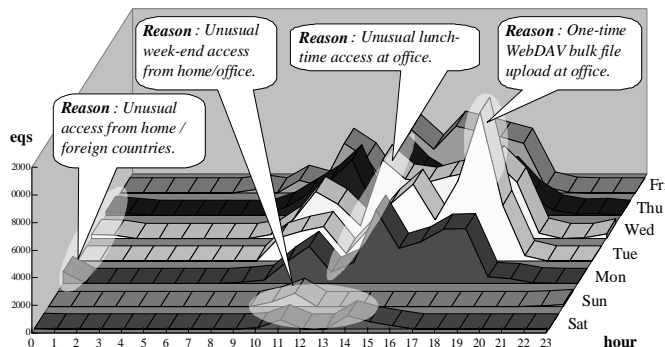


図2 User₁の2007年度のサービスアクセス傾向

図2は、ある User₁ の 2007 年度一年に渡る当該ウェブサイトへのアクセス頻度を時間・曜日別に示したものである。このユーザーは、通常平日の通常業務時間に多くサービスを利用し、昼休みを規則的にとる傾向がみられる。そこで、提案処理方式における日常性の判定計算にこれらを用いることが可能である。

一方、頻度が少ないアクセスについても、事前に準備しておいた理由づけデータベースのヒューリスティック規則により、その理由を自動判別できる可能性が見て取れる。図中では、まれな自宅や外国からの深夜のアクセス、自宅やオフィスからの週末のアクセス、昼休みにおけるアクセス、一度だけしか行っていない大量の WebDAV によるファイルの更新などである。

いずれも、日常とは違う状況におかれたユーザーの状況がアクセス時に判別できれば、それぞれに適切なサービスを適切な品質や優先度で提供することできる有用な場面だと考えられる。

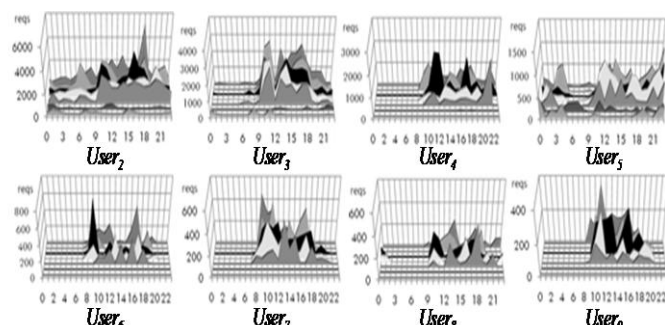


図3 複数ユーザーの2007年度のサービスアクセス傾向

図3には、他の複数名のユーザーの事例を示す。ここでは、利用者のサービス利用頻度によらず、各ユーザー別に異なるアクセスパターンの大きな違いに注目できる。例えば、User₂ は RSS リーダーによる自動アクセスを行っており、これらのアクセスを排除すると実ユーザーによるアクセスが判別できる。User₃ は 出社前に自宅からアクセスを行い、その後の通勤時間におけるアクセス停止が見られる。そこで出社時間を判定するヒューリスティック規則が検討できる。その他、初めての数回のアクセスの検出規則なども考えられる。これは、初心者がサービスにアクセスしてきた際に手厚いサポートを提供すること等に利用できる。

4. まとめ

ユーザー行動データの自動処理方式を提案し、利用法についての検討を行った。今後は、実サービスへの適用を図りたい。

参考文献

[宮内 09] 宮内興治: 企業情報共有システムに見られたユーザーの型, 知識流通ネットワーク研究会 第四回研究会, 人工知能学会, 2009.