3E3-OS20-1in

Web上のライフストリームの構造化に関する考察

On Structuring Lifestream Data on the Web

長野 伸一[*1] 佐々木 健太[*1] 上野 晃嗣[*1] 長 健太[*1] 川村 隆浩[*1]
Shinichi Nagano Kenta Sasaki Koji Ueno Kenta Cho Takahiro Kawamura

[*1](株) 東芝 研究開発センター
Corporate R&D Center, Toshiba Corporation

A lifestream is the stream data of human daily activities, which include both physically-sensed data in the real world and personal activity data on the Web. In integrating stream data from heterogeneous sources, it is essential to give semantics to elements of the lifestream. Linked Data is the collection of hyperlinked data on the Web. Applying Linked Data as semantic resources is one of the solutions for this purpose. The paper surveys the related work on sensing and capturing human activities both from the physical and virtual worlds, and then discusses the issues in exploiting Linked Data for structuring lifestream.

1. はじめに

ライフストリームとは，ユーザの行動に関して Web 上を流れる情報の経路をいう [Freeman96]．一般に，ユーザは利用目的に応じて，Web 上で提供される複数のパーソナルサービスやソーシャルサービスを使い分けている．ユーザの行動支援を目的としたサービスの実現という観点からすれば，そうした複数のサービス上に蓄積されたユーザ毎のデータを 1 本のストリームに集約することが望ましい．一方，ヒトが携帯または装着できる小型サイズのセンサデバイスや，据え置き型のインフラセンサが登場し始め，実世界におけるユーザの行動や状況に関するセンサ情報までもが Web 上で流通し始めている．ユーザ行動に関するこれらのオンライン情報とオフライン情報とを集約し，ストリーム型のデータとして扱うことにより，ユーザの行動や関心を途切れることなく追跡し，把握する新たな技術が現実のものへとなりつつある．

こうした異なる複数の情報源からのストリーム型データを集約するにあたり，データを構造化して意味を与え，データ間どうしを関連づけることが不可欠となる．実世界情報への意味づけを行うには，Linked Open Data と呼ばれるオープンな Web データは有用な資源の 1 つと言えよう．本稿では，関連する取り組みを俯瞰することにより，ライフストリームデータの構造化に向けた課題について議論する．

2. 実世界情報の Web への公開

実世界で活動するユーザから，行動や感心に関する情報を取得するには，何らかのセンサが必要である．センサは，その設置場所によって大きく，装着型と環境型とに分類される．

装着型は，小型のセンサ機器をユーザが体に身につけたり携帯するタイプのものである．スマートフォンに代表される，精度の高いセンサ機器を搭載したモバイル機器の普及により，ヒトの行動を計測するライフログデバイスとしての活用が進みつつある．KDDI が取り組むサービス「ケータイ de ライフログ」は，携帯電話に搭載のセンサ機器を用いて，利用者の身の回りの情報を取得し，サーバに収集し，ネットワーク上にある情報と合わせて，いつ・どこで・誰と・何をしたのか，何に興味を持ったかなどの情報がライフログとして記録している [小塚 09]．ライフログ情報の管理には，セマンティック Web の記述言語 RDF を採用し，3 つ組を基本構造とするグラフデータとして格納している．一方，オフィスでは業務に PC を利用することが多い．そこで，PC の操作履歴を収集するソフトウェアをインストールし，Web 閲覧，メール送受信などの履歴をもとに，ユーザの行動を獲得する研究が進められている．Cheyer らは，タスク管理や時間管理の支援を目的として，セマンティックデスクトップのシステム SRI を開発し，膨大な操作履歴を分析し，ユーザの行動や意図を推定している [Cheyer05]．

一方，環境型は，実世界にセンサ装置を設置するもので，周辺環境を観測したり，不特定多数のユーザの情報を取得する．Microsoft Research の SenseWeb[*1]はその代表例で，世界中に設置したセンサータ機器から周辺環境のリアルタイム情報を収集し，Web 上で提供している．センサから得られるストリームデータをもとに，ヒトの行動を認識するには，大量のデータを収集し分析し，ストリームデータに対する行動のラベリングを行う必要があるが，大量のデータを収集するには時間もコストもかかる．人間行動理解に関する研究開発を加速する目的として，実空間情報を集めて公開している取り組みが進められており，例えば，MIT の Place Lab[*2]，人間行動センシングコンソーシアム (HASC)[*3]などがある．

3. Web からの実世界情報の抽出

ブログや Twitter などのソーシャルメディアは，ユーザが体験や意見を発信し，他のユーザとリアルタイムに共有する，新たなメディアとなっている．こうしたメディアは，実世界の情報とともにユーザの体験や意見が綴られる傾向にあり，オンラインだけでなくオフラインにおいてもユーザどうしが交流を深める場として広く活用されている．ソーシャルメディアを対象とした行動分析の研究の一環として，ユーザの行動に関連する実世界情報を抽出する取り組みが進められている．

代表的な実世界情報として，モノ情報と地理情報がある．モノ情報とは，行動の対象物に関する情報を指す．ユーザがソーシャルメディアを利用して発信したテキスト情報から，日常生

連絡先: 長野 伸一，shinichi3.nagan@toshiba.co.jp

∗1 http://research.microsoft.com/en-us/projects/senseweb/
∗2 http://architecture.mit.edu/house_n/placelab.html
∗3 http://hasc.jp/

活で見聞したモノに関する体験や意見などのクチコミ情報を抽出する技術は，クチコミ抽出や意見マイニングなどと呼ばれる．一般に，クチコミ抽出では，モノの名前とともに，そのモノに対するユーザの評価観点 (価格，機能など)，評価意見 (高い，使いやすいなど) を対にして抽出する．実世界のセンサを利用して，ユーザが見聞したモノを認識するには精度や実現コストの面で課題が多く，テキスト情報の活用が有効である．

一方，地理情報は，物理的な位置や建物などを指す．ユーザが発信したテキスト情報に対して自然言語処理を適用し，ユーザが実際に現地に訪問した地名を抽出できれば，ユーザの滞在地や動線の分析が可能となる．単純には，地名辞書を利用して文章中から地名を抽出できる．例えば,「東京」は通常，地名として認識されるであろう．しかし，実際には「東京」について言及していないものや,「東京」に滞在していないものもある．松本 [松本 10] らは，地名と共起する単語，および地名の位置情報とを併用して，主題となる地名を効率よく抽出する方式を提案している．

4. Linked Data の活用

テキスト情報からモノ情報，地理情報を抽出し，GPS などの実世界情報と付き合わせ，個人行動情報の意味づけを行うことにより，行動プロファイリングなどの分析が可能となる．さらには，行動情報を論理的に集約することにより，モノや場所を介したヒトどうしのつながりが可視化され，新たなソーシャルサービスの実現が可能となるであろう．

このような行動情報の意味づけを行うにあたり，Web 上のオープンなデータを活用することは自然な流れである．近年 Linked Data と呼ばれる，Web データの公開が進められている．Linked Data は，個々のデータが URL を持ち，データ間にハイパーリンクが付与された，データの Web である．Linked Data の中には実世界におけるモノ，コトに関するデータも整備されつつあり，Linked Data を概念辞書と見なすことにより，ライフストリームデータへ意味情報を与えることが可能となる．Zander らは，図 1 に示すように，モバイル機器を対象として，機器から獲得される情報を Web 情報とを集約するアーキテクチャを提案している [Zander10].

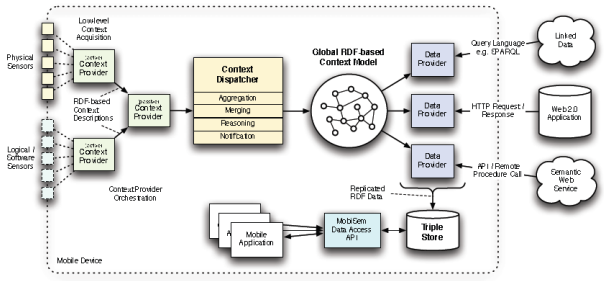

図 1: 実世界情報とオンライン情報とを集約する，モバイル機器向けアーキテクチャの例 [Zander10]

一般に，ユーザ行動を関連する属性情報としては，時間，場所 (地名や施設など)，対象物 (商品や催しものなど)，ヒト (個人や組織など)，手段 (交通や道具など) が挙げられる．例えば，場所に関する Linked Data には，LinkedGeoData，GeoNames などがある．一方，対象物はドメインに依存する．DBpedia のようなハブには著名なモノ情報はあるものの，一般にはドメインごとのデータを利用するのがよく，例えば，音楽情報であれば MusicBrainz を利用できる．また，Linked Data ではないものの，製品情報を記述するマイクロフォーマット hProduct が，検索エンジンのスニペットや小売オンラインサイトで利用され始めている．

一方，ライフストリーム自体を記述するデータモデルも必要であろう．関連するモデルの 1 つとして，Event Ontology がある．Event Ontology はイベント (事象) を表現するモデルで，1 つのイベント (事象) は，agent, time, place, factor, product の 5 つのプロパティを持つ．また，ヒトどうしのソーシャルな関係を表現するモデルとして，FOAF が広く利用されている．

今後，ライフストリームデータの集約，活用に向けて，解決すべき課題としては以下のものが挙げられる．

- ライフストリームを表現するデータモデルの標準が不可欠である． 既に Web 上で利用されている Linked Data や語彙との親和性に配慮した設計が求められる．
- 一般に，実世界のセンサ機器から獲得されるデータはプリミティブなものが多い．ヒトを主体とするライフストリームの粒度で扱えるよう，ライフストリーム集約の枠組みの中でセンサ情報を抽象化もしくは解釈する仕組みが必要である．
- ライフストリームを Linked Data と融合することにより，Linked Data のクエリ言語 SPARQL を利用してストリームデータの検索が可能となる．SPARQL は，時間情報や空間情報のように連続性のあるデータを扱いづらく，拡張が求められる．

5. おわりに

ヒトの行動に関連する実世界情報とオンライン情報とが集約されたライフストリームデータを扱うにあたり，ライフストリームのデータ構造化と意味付けに関連する取り組みを俯瞰し，解決すべき課題について述べた．今後は，ライフストリームのデータモデルを検討していく．

参考文献

[Freeman96] Eric Freeman and David Gelernter: Lifestreams: A Storage Model for Personal Data, ACM SIGMOD Bulletin (1996).

[松本 10] 松本光弘, 二宮亜佐美, 長岡涼, 沼尾正行, 栗原聡: WWW 上の文章に含まれるイベント情報からの地名の同定, 第 88 回知識ベース研究会 (SIG-KBS) (2010).

[Chen10] Hsinchun Chen and David Zimbra: AI and Opinion Mining, IEEE Intelligent Systems, Vol.25, Issue 3, pp.74-80 (2010).

[Zander10] Stefan Zander, Bernhard Schandl: Context-driven RDF Data Replication on Mobile Devices, Semantic Web Journal (2010).

[小塚 09] 小塚宣秀, 森川大輔: ケータイ・ライフログとしての実空間プロファイルと流通・管理技術, 情報処理, Vol.70, No.7 (2009).

[Cheyer05] A. Cheyer, J. Park, R. Giuli: IRIS: Integrate. Relate. Infer. Share. 1st Workshop on The Semantic Desktop (2005).