2F2-2in

# オフィスワーク向けライフログ生成のための センサデータと操作履歴からのキーワード抽出

## Term Extraction from Sensor Data and Operation History for Office-work Lifelog

岡本 昌之[*1] Masayuki Okamoto　渡辺 奈夕子[*1] Nayuko Watanabe　長野 伸一[*1] Shinichi Nagano　長 健太[*1] Kenta Cho　川村 隆浩[*1] Takahiro Kawamura

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

One of largest problems of lifelog-based application is readability of the stored lifelog though the number of lifelog platforms is increasing. We propose a term extraction method to add annotation labels to the stored lifelog for supporting office worker, exploiting text data acquired from desktop activities. Our prototype system monitors a user's desktop activities after combining raw events, then extracts possible annotation labels with LDA and C-value techniques from documents and text data in sensor events.

## 1. はじめに

PC や携帯電話・スマートフォンなど個人で利用する端末にセンサデバイスが多数登載されるのに伴い，日常活動記録であるライフログが蓄積されるようになりつつある．しかしながら，蓄積されるデータが急増する一方，そのデータを個人が活用するシーンはまだ少ないと言える．理由として，データそれ自体をユーザが後から活用したいと考える場面が少ないこと，およびデータを簡単に振り返る手段が提供されていないことが挙げられる．

我々は，まず後者に注目し，ログに対し適切なアノテーションを付与することで簡単に検索・閲覧する枠組み作りを目指している．有用なドメインの 1 つにオフィスワークが挙げられるが，既存の研究ではユーザの活動履歴に応じたアノテーションの自動付与はほとんど行われていない．また，後で検索するためのアノテーションの要件として，トピックを分類できること，可読性の高い単語であること，の 2 点が挙げられる．

これまで，情報抽出分野では，様々なドキュメントからキーワードを抽出する手法が提案されている．文書閲覧履歴からのキーワード抽出では，Web 閲覧履歴からの重要語抽出が行われている[Matsuo 2002]．また，作業履歴から資料を分類するアプリケーションとして，TaskPredictor2 が提案されている[Shen 2009]．しかし，後から行動履歴としての文書閲覧履歴を簡単に振り返るために可読性の高いアノテーションを付与する方法についての議論はほとんど行われていない．

本稿では，オフィスワークを対象としたライフログ記録モジュールの紹介と，記録されたライフログへのアノテーション方式の 1 つとして，キーワード抽出に基づくライフログへのアノテーション付与について述べる．

## 2. オフィスワーク向けライフログ生成

ライフログ生成の流れを図 1 に示す．本稿では，オフィスワークで一般的に用いられる PC における記録を想定する．まず，各種センサやアプリケーションからのイベントを記録する．記録ツールは Microsoft Windows 上の常駐アプリケーションとして動作し，操作情報としては Microsoft Office での利用ファイル，Web ブラウザ（Internet Explorer, Mozilla Firefox, Google Chrome）における閲覧 URL およびタイトル，Outlook における送受信メールのタイトル，その他操作中のアプリケーションのプロセス名およびウィンドウタイトルが記録される．また，WiFi，マイク，Bluetooth などのデバイス状態に関してはポーリングにより定期的に状態を記録する．各イベントはイベントバッファに入力され，順に記録対象の行動であるか判定するイベント判定処理が実行される．イベント判定では，一定期間内に入力されたイベントの組み合わせに対し，時間幅を持った行動エントリ（例えば，メール作成開始イベント発生からメール送信イベント発生までの間をメール作成行動として記録）を生成してデータベースに格納する．あわせて，付随するテキスト情報はコンテンツとして記録される．一定期間毎に，蓄積されたライフログに対しコンテンツから抽出されたキーワードがアノテーションとして付与される．

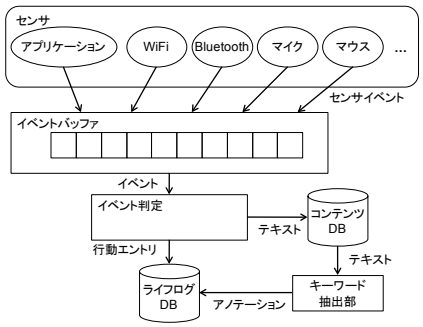


図 1 ライフログ生成処理の流れ

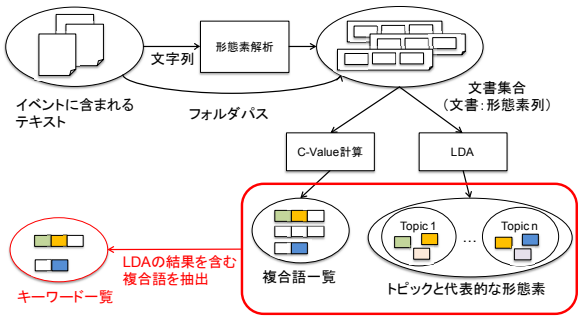


図 2 キーワード抽出の流れ

## 3. キーワード抽出

得られたログのうち，テキストデータを含むライフログ情報から，アノテーションに用いるキーワードを抽出する．キーワード抽出

連絡先：岡本昌之，(株)東芝 研究開発センター，〒212-8582 川崎市幸区小向東芝町 1，
masayuki4.okamoto@toshiba.co.jp

の流れを図 2 に示す．まず，操作イベントからテキスト情報を抽出する．テキスト情報には，ウィンドウタイトルや URL，利用ファイル名，ファイルに含まれるテキスト情報が含まれる．次に，これらのテキストデータから適当なキーワードを抽出する．オフィス文書の主な特徴として，各文書は特定のトピックに属することが多いことが挙げられる．したがって，トピック分類とトピック毎の重要語を抽出することが望ましいと考えられる．また，処理対象の期間を変えることで，1 日の中で詳細なトピックを検索したり，長期間の中でより後半なトピックを俯瞰する枠組みを構築できるようになる．

本稿では，トピックを表す主要な単語の抽出に LDA (Latent Dirichlet Allocation)を，可読性の高い複合語の抽出に C-value 法を用い，組み合わせることでアノテーションに利用可能なキーワードを推定する手法について述べる．

### (1) 前処理

キーワード抽出の前処理として，処理対象の期間に含まれる各文書に対し形態素解析を行い，形態素列に変換する．

### (2) LDA によるトピック推定・重要語抽出

LDA (Latent Dirichlet Allocation) [Blei 2003]は，与えられた D 個の文書集合を K 個の潜在トピックの混合とみなし，また各潜在トピックは W 種類の単語の多項分布から構成されるとみなす生成モデルである．潜在トピックを生成するモデルのパラメタ $\theta$を Dirichlet 分布Dir(α)で表し，文書 $j$ に対する$N_j$個の単語について，トピック $z_{ij} = \mathrm{Multinomial}(\theta_j)$，単語 $x_{ij} = \mathrm{Multinomial}(\phi_{z_{ij}})$でモデル化するものである．本稿では，ライフログとして記録された文書集合を入力としてトピック推定およびキーワード抽出を行う．パラメタ推定には，Fast Collapsed Gibbs Sampling [Porteous 2008] を用い，100 回のイテレーション結果のうちパープレキシティが最小となるトピック数を採用し，この中でトピック毎の出現確率が高い単語を候補とする．

### (3) C-value を用いた複合語の抽出

形態素そのままでは，ユーザが解釈するには形態素単位では細かすぎると考えられる．複合語抽出の手法として，C-value [Frantzi 1996]を用いる．複合語 $a$ に対する C-value は，$a$ を含むより長い複合語の頻度を $t(a)$，候補数を $c(a)$として

$$Cvalue(a) = (length(a) - 1)\left(freq(a) - \frac{t(a)}{c(a)}\right)$$

と表される．C-value が高く，かつステップ(2)で抽出された単語との重なりが大きい語を，アノテーション候補として抽出する．

### (4) キーワードの選定

前述のステップ(2)，(3)にて得られたそれぞれの単語集合を比較し，C-value が高い複合語のうち，LDA によるトピック毎のお生成確率の高い形態素を含む単語をキーワードとして採用する．

実際の業務記録を対象にアノテーション候補を抽出した例を表 1 に示す．ここでは，402 イベントを対象としたもので，調査のトピックや研究テーマ，論文執筆作業等に関する資料の確認作業が含まれる．「PowerPoint のタイトル」のように，文書に含まれるがアノテーションとしては意味のない語の除外は今後の課題である．また，文書数およびトピック数を変化させた時の関係をパープレキシティの推移として表示した結果を図集合全体に対し C-value を算出した結果の上位 10 件を図 3 に示す．Gibbs サンプリングのイテレーション数は 100 回に固定し，横軸はトピック数を，縦軸はパープレキシティを表す．個人が読む文書を対象とする場合，比較的処理対象の規模が小さいためイテレーション数，トピック数とも数十回程度でパープレキシティは最小となることが分かる．

**表 1** LDA により出力された生起確率が高い語との重なりが大きい C-value 上位の複合語例（赤字は重複箇所）

| |
|---|
| 特集一覧，テーマ・キーワード，グループ 18，今後のライフログアプリ開発方針，企業の会議・コミュニケーション支援技術に関する市場動向，キーワード抽出の流れ，ライフログ活用 11 上，PowerPoint のタイトル，各ドキュメントから抽出している属性，行動収集進捗，ライフログアプリ |

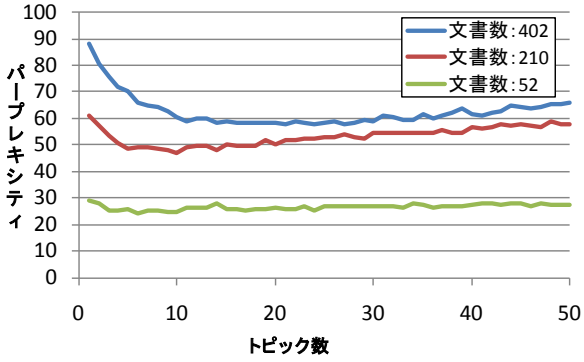


**図 3** 文書数とトピック数，パープレキシティの関係

## 4. おわりに

本稿では，オフィスワークを対象としたライフログ記録と，記録されたライフログに対するアノテーション付与を想定したキーワード抽出方式について紹介するとともに，実際の業務文書を対象とした処理結果の例を紹介した．本稿ではキーワード抽出自体に関して述べたが，付与すべきラベルやトピックに関しては時間とともに変遷がある場合も多い．また，本稿では個人の履歴のみに注目したが，他人のセンサ情報やライフログを考慮することで，例えばコミュニケーション記録[Okamoto 2008]もライフログとして活用可能である．

## 参考文献

[Blei 2003] D. M. Blei et al.: Latent Dirichlet Allocation, *J. Machine Learning Research*, Vol.3, pp.993-1022, 2003.

[Frantzi 1996] K. Frantzi and S. Ananiadou: Extracting Nested Collocations, *Proc. COLING-96*, pp.41-46, ICCL, 1996.

[Matsuo 2002] Y. Matsuo et al.: Browsing Support by Highlighting Keywords based on a User's Browsed History, *Proc. SMC-02*, 2002.

[Okamoto 2008] M. Okamoto et al.: Finding Two-level Interpersonal Context: Proximity and Conversation Detection from Personal Audio Feature Data, *Proc. Interspeech-08*, pp.2482-2485, 2008.

[Porteous 2008] I. Porteous et al.: Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation, *Proc. KDD-08*, ACM Press, 2008.

[Shen 2009] J. Shen et al.: Detecting and Correcting User Activity Switches: Algorithms and Interfaces, *Proc. IUI-09*, pp.117-126, 2009.