Wikipedia 記事の内容と閲覧時間帯の関係の統計的分析

吉井 健敏 [1] 持橋 大地 [2]
[1] 株式会社 D2C [2] 統計数理研究所
taketoshi.yoshii@d2c.co.jp

概要

インターネット上には様々なコンテンツが溢れていて、ユーザはライフサイクルに沿ってそれらを利用する。メディアは閲覧数に応じた収益を獲得するため、コンテンツとユーザの時間特性を把握し閲覧数を増大させることは重要なビジネステーマである。本研究では Wikipedia の記事内容と閲覧ログを用いて、内容とユーザの興味の時間特性を明らかにする手法を提案する。最初に内容が類似する Wikipedia 記事は閲覧される時間帯が類似することを示す。次に独立成分分析を利用して異なるユーザグループの時間特性を抽出し、それらが直感的なライフサイクルと一致することを確認する。これまでの議論から記事内容と閲覧傾向に密接な関係があることが期待できるため、最後にニューラルネットを使って記事内容から閲覧傾向を直接予測できることを示す。本研究は Wikipedia の記事について注目しているがシンプルな手法ゆえ、閲覧ログが取得可能なあらゆるコンテンツで適応可能である。

1 はじめに

インターネット上にはブログやニュース、動画など多岐にわたるコンテンツが溢れている。メディアを運営する企業はそれらの閲覧数に応じた収益を獲得するため、閲覧数を最大化する戦略を模索することは重要なビジネステーマである。一方でユーザは自身の生活サイクルや興味に則ってインターネットを利用し、この営みが Web 上のコンテンツの時間特性に影響を与えている。このような時間特性を解明することは、コンテンツの内容とその配信タイミングを最適化する戦略に大きく貢献することができる。

本研究では、Wikipedia の記事閲覧ログを用いて、記事内容とユーザの時間特性の関係を明らかにする手法を提案する。まずはじめに Wikipedia データを定量的に扱うため、記事内容と閲覧数の時間変化を多変量へと変換する § 3.1。その後 K-means を用いて、それぞれの空間内で内容クラスタと閲覧クラスタを定義し、それぞれの特性を調査する。手始めに同一の内容クラスタに含まれる記事が似た様な時間特性を示すことを確認する § 5.1。次に閲覧の時間特性がユーザのライフサイクルや時間に沿った興味によって生じていることを明らかにし § 5.2、似た様な時間特性を示す記事は内容も類似することを確認する appendix A.1。最後に内容ベクトルから閲覧ベクトルへの変換がニューラルネットを用いることで可能であることを示すことで § 5.3、記事内容とユーザの時間特性に密接な関係があることを示す。本研究は Wikipedia の記事に注目しているが、閲覧ログを取得可能なあらゆるインターネット上のコンテンツに適応可能で、目的の配信効果を狙ったコンテンツの編集にも応用することが期待できる。

2 予備知識

2.1 PMI 行列の行列分解を用いた Doc2Vec

Word2Vec [1] や Doc2Vec [2] は単語や文章を線型空間に埋め込み、多変量として扱うための手法だが、ニューラルネットを使用するため計算コストが高くなってしまう問題点があった。Levy et al. 2014 [3] では、行列分解を用いて Word2Vec や Doc2Vec と等価な計算を行うことで、文章と単語を同時に線型空間内に埋め組む手法を提案している。複数の文章で構成されたデータセット $S = \{s_i\}$ について、文章 s ごとに単語 w の登場頻度 $n(w,s)$ を集計し、

$$\mathrm{PPMI}(w,s) = \max\Big(\log \frac{P(w,s)}{P(w)P(s)}, 0\Big) \qquad (1)$$

を計算する (ただし、$P(\cdot) = n(\cdot)/\sum n(\cdot)$)。式 (1) から得られる行列 $M = (\mathrm{PPMI}(w,s))$ について特異値分解 $M = U\Sigma V^\top$ をして得られる行列 $U\sqrt{\Sigma}$ は Word2Vec と等価な結果が、さらに $V\sqrt{\Sigma}$ からは単語ベクトルと同一空間内に埋め込まれた文章ベクトルを獲得できる。この手法の優れた点は、スパースな行列分解を

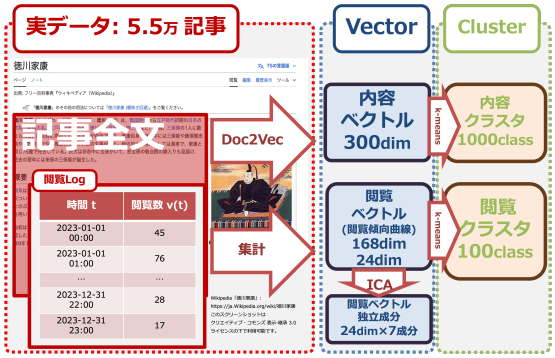

図 1: 研究で扱うデータとそれら変換方法の概略図。

使用するため計算コストが非常に軽量で済むことである。

2.2 ICA (独立成分分析)

独立成分分析 (ICA) は観測で得られた N 個の T 次元のベクトル $X = (\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_N)^\top \in \mathbb{R}^{N\times T}$ を M 個の独立なベクトル $S = (\mathbf{s}_0, \mathbf{s}_1, \cdots, \mathbf{s}_M)^\top \in \mathbb{R}^{M\times T}$ の線形和で表現する手法である。混合行列を $A \in \mathbb{R}^{N\times M}$ とすると

$$X = AS \tag{2}$$

と表すことができ、A は $M \leq N$ ならば解が存在する。ここで $\mathbf{s}_i$ に非正規分布を仮定し、$\mathbf{s}_i$ が独立という条件

$$p(S) = \prod_{i\leq M} p(\mathbf{s}_i) \tag{3}$$

を満たしながら、$S = WX$ となる W を探索することができ [4]、混合行列 $A = W^{-1}$ が得られる。

3 提案手法

3.1 Wikipedia 記事からの情報抽出

本研究で使用する Wikipedia データでは、記事タイトル $s_i \in S$ ごとに記事内容 a_i と 1 時間ごとの閲覧数 $v_i(t)$ の情報を有する。これらの情報を定量的に扱うために次のような処理を施す。

内容ベクトルと内容クラスタ Wikipedia 記事 $s_i \in S$ について内容 a_i がある。この自然言語情報を定量的に扱うため § 2.1 で紹介した手法を使用する。まず、記事内容の数字を#に置換し、IPA 辞書を使用して MeCab [5] で形態素に分解する。記事 s ごとの形態素 w の登場頻度 $n(w, s)$ を集計し、式 (1) より行列 $M = (\mathrm{PPMI}(w, s))$ を作り、特異値分解 $M = U\Sigma V^\top$ して得られる行列 $V\sqrt{\Sigma}$ から記事内容のベクトル表現を作成する。本実験では特異値分解のランクを 300 として、記事ごとに 300 次元の内容ベクトルを獲得する。こうして得られた内容ベクトルについて、K-means を用いて 1,000 クラスに分割し、それらを内容クラスタとして扱う図 1。この内容クラスタは単語の登場頻度が類似した記事が同一のクラスタに属する。

閲覧ベクトルと閲覧クラスタ 次に記事が閲覧される時間帯の傾向を示す表現を作成する。Wikipedia 記事は記事ごとに閲覧数が桁で異なるため、規格化 $v'_{i,T} = \log_{10}(v_{i,T}/\bar{v}_i)$ し、時間帯別の変動を捉えやすくする。時間帯 T は集計方法により異なるが、本研究では時間別 (0、1、...、23 時台; 24 要素)、曜日時間帯別 (月 0 時、...、日 23 時; $24\times 7 = 168$ 要素) をそれぞれ採用している。これらを要素とするベクトル $(v'_T)_i$ を記事ごとの閲覧ベクトルとする。このとき 168 次元の閲覧ベクトルに対して K-means を用いて 100 クラスに分割したものを閲覧クラスタと呼び、似たような時間帯に閲覧される記事が同一のクラスタに属する。

3.2 独立成分分析を使った時間帯別閲覧傾向の抽出

Wikipedia には様々な内容の記事が投稿されていて、閲覧される時間はユーザの生活サイクルや時間帯に沿った興味によって決定される。ユーザの生活サイクルや興味とは例えば、数学や歴史の記事は学生の調べ物に利用されるため平日の日中に閲覧されやすかったり、深夜アニメに関する記事は休日前の夜中に閲覧されやすかったりする傾向のことで、それらには大局的なグループが存在し、各グループはそれぞれ異なる時間的振る舞いを示す。これらグループの重ね合わせによって、記事の時間帯別閲覧数が記述されるため、独立成分分析によりそのグループごとの時間特性を抽出する。記事の時間帯別閲覧数 $v'_i(T)$ からユーザグループごとの閲覧傾向を抽出するため、閲覧ログに独立成分分析を実施する。記事ごとの閲覧ベクトル $V' = (v'_{i,T}) \in \mathbb{R}^{N\times T}$ について独立成分分析で $S = WX'$ となる独立成分 S を求める。このとき抽出する成分数を M とすると $S \in \mathbb{R}^{M\times T}$ となり、独立成分を時間の関数 $S(t|m)(m \in \{1, \cdots, M\})$ と見なすことで、独立したユーザグループ (m) の閲覧傾向の時間変化を導く。

4 実験データ

4.1 Wikipedia データ

Wikipedia に関する 2 つの公開データを使用する。**時間帯別閲覧数ログ**[1] 一つ目は 2015 年 5 月以降のすべての Wikimedia 財団プロジェクトの閲覧数がまとめられたデータセットで、1 時間ごとに閲覧された回数 $v_{i,t}$ が記事ごとにまとめられている。本研究では季節性の影響をなくすため 2023-01-01~2023-12-31 までの 1 年間のログを使用する。また分析する記事を日本語記事に限定することで、閲覧の大半が日本からのものに絞ることができ、日本時間での活動の傾向を抽出することを狙う。

日本語記事[2] 2 つ目は Wikipedia の日本語記事の内容をまとめたデータセットで、2024-01-01 時点で約 140 万件の記事が収録されている。

4.2 記事のスクリーニング

本研究では Wikipedia にアクセスするユーザのライフサイクルと関心の時間遷移を明らかにすることを目的としている。そのため Wikipedia データのすべての記事を分析する必要はなく、不要なタイトルを取り除く処理を行う。取り除く基準は以下の 2 つで、(1) 極端に閲覧数が少ない記事タイトルの除外。記事ごとの年間閲覧数は桁で異なり、中には年間通してほとんど閲覧されないタイトルも存在する。それら記事はノイズとなるので年間閲覧数が 10,000 件以下のものは取り除く (記事数: 140 万件 →6.1 万件)。(2) 閲覧数が短時間で劇的に変化したものを除外。Wikipedia 記事の中には事件やニュースなどで突発的かつ劇的に閲覧数が上昇するようなものが存在する。これらは日々のライフサイクルに注目する上でノイズとなるので、年間を通して閲覧数が 1 時間のうちに 100 倍上昇/下降したことがある記事を除外する (記事数: 6.1 万件 →5.5 万件)。

5 結果

5.1 内容クラスタごとの閲覧傾向曲線

§ 3.1 で定義する内容クラスタが、それぞれどの時間帯に閲覧されやすいのかを示す閲覧傾向曲線を作成する。一つの内容クラスタ (CC; Context Cluster) に含まれる Wikipedia 記事の総閲覧数の時間変化 $v(t|CC) = \sum_{i \in CC} v_i(t)$ について曜日時間帯 T ごとに集計 $V_T^{CC} = \sum_{t \in T} v(t|CC)$ し規格化 $V_T'^{CC} = \log_{10}(V_T^{CC}/\bar{V}'^{CC})$ したものが閲覧傾向曲線 図 2(左) である。この閲覧傾向曲線についてすべての内容クラスタについて日曜の変動を描画したものが 図 2(右) で、深夜から早朝にかけて閲覧数が減少し、昼ごろから夜にかけて閲覧されやすくなるという、一般的な休日の生活リズムをよく表している。一方で日曜の 20 時頃にいくつかの内容クラスタが全体の傾向から逸脱していることがわかる。これはその内容クラスタに含まれる記事が他の記事よりも同時間帯で閲覧されやすくなっていることを示している。日曜日の 20 時台に閲覧されやすい内容クラスタの Top4 とそれに属する記事をまとめたものが 表 1 で、内容が戦国時代から江戸時代にかけての歴史上の人物・用語が多く登場していることがわかる。これは 2023 年の大河ドラマが徳川家康をテーマにしたもので、大河ドラマの放送終了直後からそれに関連した人物やものが調べられた結果と推測できる。これらの結果から同一の内容クラスタは似たような時間帯に閲覧されやすいことが示唆できる。

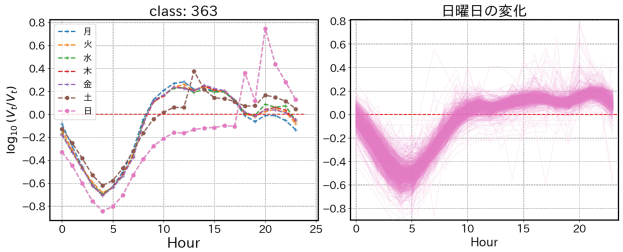

図 2: (左) 内容クラスタ (CC=363) について閲覧傾向を可視化したもの。(右) 全内容クラスタについて、日曜日の閲覧傾向を可視化したもの。

表 1: 日曜 20 時台に人気の内容クラスタと属する記事一覧

Class363	Class166	Class802	Class152
徳川家康	本多忠勝	小堀政一	後藤寿庵
織田信秀	榊原康政	於大の方	織田信長
小早川秀秋	織田秀信	徳川四天王	毛利良勝
楽市・楽座	結城秀康	徳川秀忠	服部一忠
顕如	宗義智	豊臣秀頼	武田勝頼
...	...	...	...

1) https://dumps.wikimedia.org/other/pageviews/readme.html
2) https://dumps.wikimedia.org/jawiki/

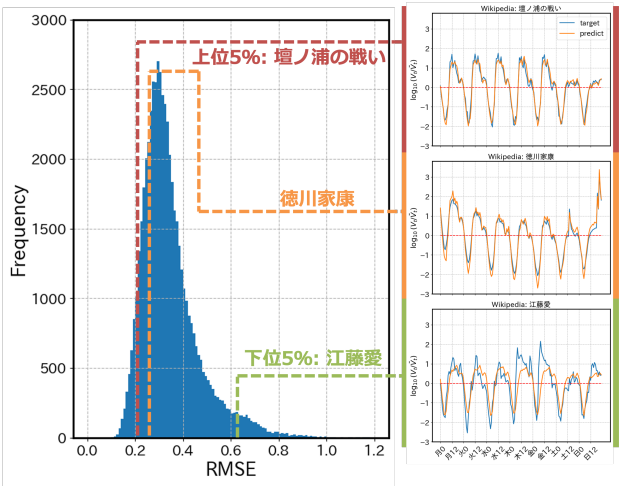

図 3: ニューラルネットの推論とターゲットの比較

5.2 ICA を使った時間帯別閲覧傾向の抽出

§ 3.1 に従って時間ごとに集計した 24 次元の閲覧ベクトルについて成分数を 7 とした独立成分分析を実施し、成分ごとの時間特性を可視化したものが 図 5 である。成分の順番は成分の強度分布の歪度が降順となるように決めていて、より特徴的な成分が若い番号となっている。各記事の閲覧ベクトルはこれら独立成分の線形和で表現でき、独立成分はユーザの閲覧傾向を示している。独立成分の極性に特定の意味はなく、同一成分内で極性が反転していればその時間帯での閲覧されやすさが逆転することを意味している。例えば成分 1 に注目すると、成分 1 の weight が正の記事については 13 時ごろに閲覧されやすく、18 時ごろに閲覧されづらい特性を持ち、逆に weight が負の記事については 13 時ごろに閲覧されづらく、18 時ごろに閲覧されやすい記事であることが把握できる。

成分 1~3 では特定のピークがあり、これらはお昼休憩や午前中の仕事時間による傾向を、成分 4~7 では日中、夜間、深夜帯などの広い時間間隔での傾向が表れていることが期待される。

5.3 内容ベクトルから閲覧ベクトルへの変換

これまでの議論および Appendix により、内容が似ている記事は似たような時間で閲覧されることが期待され、s 似たような時間帯で閲覧される記事は似たような内容の記事であることが確認できた。そこで最後に、記事内容から閲覧傾向を直接予測することができるのかを確認する。予測で使用するモデルは、入力に内容ベクトル (300 次元)、中間層 512 次元、出力に閲覧ベクトル (168 次元) となるような 2 層のニューラルネットワークを採用し平均二乗誤差 (MSE) が小さくなるように学習を進める。モデルはオーバーフィットしないように交差検証で 5 つのモデルを作成する。記事ごとの推論結果とターゲットとの RMSE の分布が 図 3 のヒストグラムである。ヒストグラムの右側には推論結果とターゲットの関係を示した 3 つの線グラフがあり、それぞれ予測精度の上位、下位 5%のものと、日曜日に特徴的な傾向を示す “徳川家康” の記事のものである。推論結果は日中が閲覧されやすく、夜中は閲覧されずらいという傾向を掴みつつ、大河ドラマ放送時間の日曜夜のピークもしっかり抑えられていることが確認できる。一方で RMSE の分布は右側にロングテイルを引いていて、 図 3 だと “江藤愛” のような特定の人物について精度が悪くなっていた。“江藤愛” は女性アナウンサーで、出演する番組の時間が閲覧される時間に強く影響しているのに対して、Wikipedia 記事から把握できる肩書きではそこまで詳細な傾向を把握できなかったことが原因として考えられる。

6 まとめ

Wikipedia の記事概要と閲覧ログを使用して、記事内容とユーザの時間特性の関係を明らかにする手法を提案した。最初に同一の内容クラスタに含まれる記事が似た様な時間特性を示すことを確認し § 5.1、閲覧の時間特性がユーザのライフサイクルや時間に沿った興味によって生じていることを明らかにし § 5.2、似た様な時間特性を示す記事は内容も類似することを確認した appendix A.1。これらを踏まえて、ニューラルネットを使用することで Wikipedia 記事の内容から閲覧傾向を直接予測できることを示し § 5.3、記事内容とユーザの時間特性に密接な関係があることを明らかにした。

本研究は Wikipedia の記事に注目しているが、閲覧ログを取得可能なあらゆるインターネット上のコンテンツで応用可能である。特にニューラルネットを利用したコンテンツ内容から閲覧傾向を予測するモデルは発表前に閲覧されやすい時間を把握できる他に、目的の配信効果を狙ったコンテンツの編集にも応用することが期待できる。

謝辞

本研究にあたって、ご討論頂いた統数研持橋先生、実験協力・論文校正を手伝っていただいた徳山氏、井上氏、大橋氏、研究の機会を与えてくださった佐野氏に感謝の意を表します。

参考文献

[1] Tomás Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. **CoRR**, Vol. abs/1310.4546, , 2013.

[2] Quoc V. Le and Tomás Mikolov. Distributed representations of sentences and documents. **CoRR**, Vol. abs/1405.4053, , 2014.

[3] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, **Advances in Neural Information Processing Systems**, Vol. 27. Curran Associates, Inc., 2014.

[4] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. **Neural Networks**, Vol. 13, No. 4, pp. 411–430, 2000.

[5] Mecab: Yet another part-of-speech and morphological analyzer. https://taku910.github.io/mecab/.

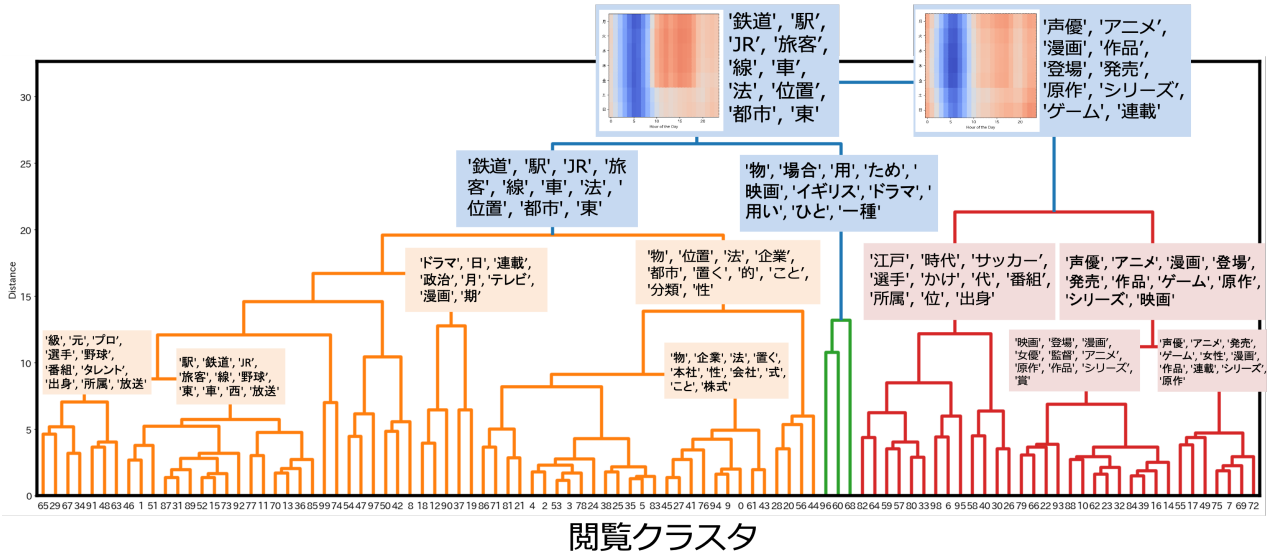

図 4: デンドログラムによる閲覧クラスタの可視化

A　参考情報

A.1　デンドログラムを用いた閲覧クラスタ構造の可視化

時間帯別閲覧傾向の独立成分の分析から、似たような時間帯に閲覧される記事は、内容も類似することが期待される。そこで、似たような時間帯に閲覧される記事で構成される 100 の閲覧クラスタ (VC; View Cluster § 3.1) を使ってその傾向を確認する。閲覧クラスタに含まれる記事について、閲覧ベクトル (曜日時間帯別 168 次元) の平均を取ることで閲覧クラスタの中心ベクトル $V_T^{\mathrm{VC}} = \frac{1}{\#(\mathrm{VC})} \sum_{i \in \mathrm{VC}} v_{i,T}$ を定義できる。このベクトルの集合について階層クラスタリングを実施することで、閲覧クラスタの構造をデンドログラムで可視化することができる 図 4。図 4 では最下層に閲覧クラスタが整列していて、距離に応じて同一視できるものが線で結ばれている。最上段の分枝にはヒートマップが掲載されているが、曜日時間帯ごとの平均閲覧数からの逸脱度 ($\log_{10}(v_t/\bar{v}_t)$) で色付けしていて、左側のクラスが平日の日中に人気で、右側が深夜帯に人気のクラスタが多く集まっていることが確認できる。線の分枝ごとに、それより下のクラスタに含まれる記事に登場する単語の中から登場回数の NPMI が高い単語を代表語として記載している。図 4 によると最上段の左側が主に交通、企業、政治のような仕事関連の記事が、右側にはアニメ漫画などの趣味に関する記事が登場することが確認できる。この構造は下に行くほど細分化されており、右側の趣味に関するクラスタに注目すると、最初にスポーツや大河ドラマ関連のクラスタとアニメ、漫画などのクラスタに分離し、アニメ、漫画関連のクラスタはさらに、映像に関するクラスタと声優に関するクラスタへと分離していく。このデンドログラムより閲覧クラスタ内では意味内容も似通ったものが集まっていることが網羅的に確認できる。

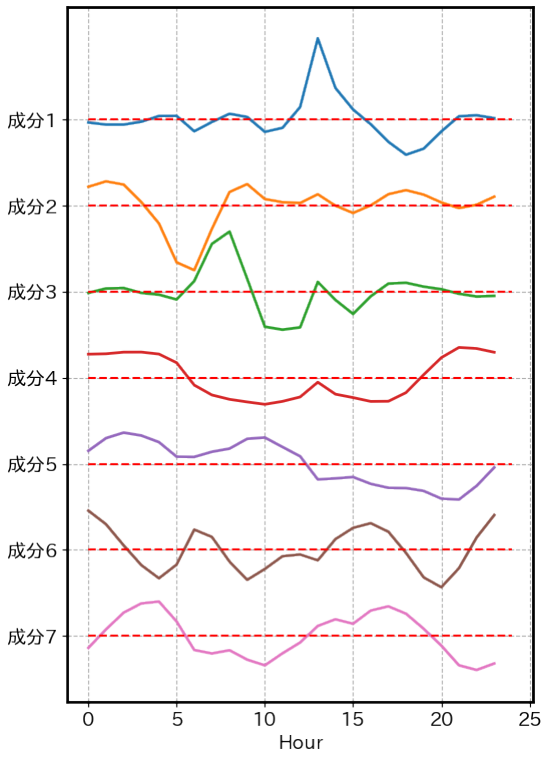

図 5: 閲覧ベクトルから得られた 7 つの独立成分