3E3-OS20-2

多項関係を活用した行為データからの興味予測

Interest Prediction from Multinomial Action Data

則　のぞみ　　ボレガラ　ダヌシカ　　石塚　満
Nozomi Nori　　Danushka Bollegala　　Mitsuru Ishizuka

We propose a method to predict users' interests, using time-evolving, multinomial relational data. We exploit various actions performed by users in social media to predict user interests. Users ' actions in social media have two fundamental properties. (a) User actions can be represented as multinomial relations - e.g. referring URLs, tagging, clicking a favorite button on a post. (b) User actions are time-varying and user-specific - each user has unique preferences that change over time. Consequently, it is appropriate to represent each user's action at some point in time as a multinomial relational data. We propose *ActionGraph*, a graph representation for modeling users' multinomial, time-varying actions. Our experimental results show that the proposed method improves the accuracy in a user interest prediction task by outperforming several baselines including standard tensor analysis, LDA-based method and several graph-based variants. Moreover, the proposed method shows robust performances in the presence of sparse data.

1. 導入

近年，人々の様々な行為に関するデータが大量に生成され，取得可能になっている．行為の例としては，様々な Social Media 上で行われている，URL（Uniform Resource Locator）やキーワードへの言及，タグ付け，お気に入り，再発信などが挙げられる．ここでは，これら人々の行為から生まれるデータを，広くアクションデータと呼ぶことにする．アクションデータは，行為の背景にある，人々の興味，嗜好を捉え，予測するのに有効であると期待できる．似たようなアイテムを好む人同士の嗜好は似ているはずだという仮説は，協調フィルタリングの基本的アイデアであるが，ここで，アイテムをアクションに拡張しても同様のことが期待できるのではないだろうか．すなわち，似たようなアクションを行う人同士の嗜好は似ているはずだ，という仮説に一定の有用性を期待できるだろう．本論文では，この仮説に基づき，人々の様々なアクションデータから，人々の嗜好，興味を予測する手法を提案する．

様々なアクションデータから人々の嗜好・興味を予測することは，推薦やパーソナライズ検索等々，様々なタスクにおいて有効であると期待できるが，この問題は比較的新しい問題であり，以下の二つの主要な課題を有すると，筆者らは考える．一つ目の課題は，アクションデータに見られる，高次元/多項関係という性質を表現し活用することである．例えば，あるユーザーが何らかのキーワードを付与してある URL への言及を行った場合，この行為には，ユーザー，URL，キーワードなどの，複数のエンティティが関わっていると見なすことができる．このように，アクションデータを扱う際には，三つ以上の共起を表現する必要が生じる．二つ目の課題は，人の嗜好、興味は時々刻々と変化していくものであり、各人の時間スケールを捉える必要があるということである．人の行為は，大まかには，長期的にも持続するような嗜好から特徴付けられると考えられるが，人は同時に，誕生日や旅行などのイベントに誘発されるような，短期間しか続かないような嗜好にも影響されると考えられる．時間を扱うモデルの多くでは，時間を，全ユーザーに共有されるグローバルな尺度として導入しているが，近い時間帯に起きた行為の中でも，同じイベントに誘発された

連絡先: nozomi.nori@gmail.com

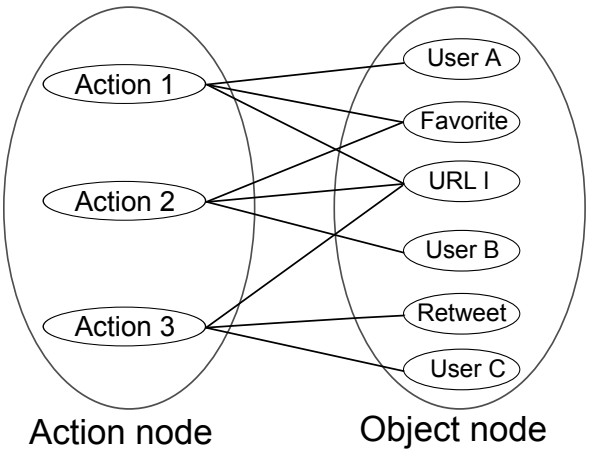

図 1: ActionGraph の表現例.

ような，偶然ではない行為というのは，特定の個人ないし少数の個人のみを含むものに限定されるだろう．従って，時間を，全ユーザーで共有されるグローバルな尺度としてではなく，各人が持っている固有の時間スケールの観点から捉える必要がある．このような，各人の時間スケールを捉え，長期的な興味に加えて短期的な興味を捉える必要性は [Xiang *et al.*2010] などで論じられており，推薦タスクにおける有用性も指摘されている．我々は，以上二点の課題を解決する手法を提案する．本論文の貢献は以下である．

- 多項関係，個々人の固有の時間スケール，という二つの性質を表現するグラフ，ActionGraph を提案する．
- 提案する ActionGraph の有用性を，実データを用いた，ユーザーの興味を予測するタスクによって評価する．LDA (latent Dirichlet allocation) [Blei *et al.*2003] ベースの協調フィルタリングや，テンソルの標準的手法である PARAFAC [Bader and Kolda2006]，その他グラフベースの様々な手法と比較し，提案手法が，予測精度とデータ過疎に対するロバスト性の観点から，有望であることを示す．

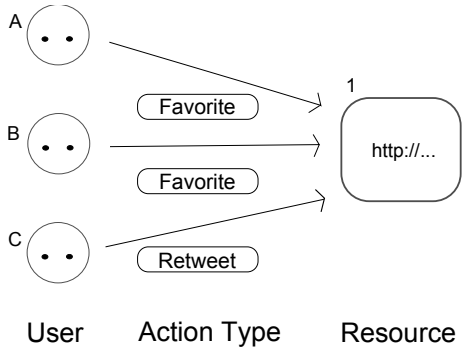

図 2: ユーザーの行為の例.

2. 提案手法

2.1 ActionGraph の構築

ユーザーによって行われる行為はしばしば複数のエンティティを含む．例えば，ユーザー A が，URL l で示されるリソースに対して，“人工知能” というキーワードでタグ付けをしたという行為を考える．この行為には，ユーザー A, URL l, 行為の動詞としての “tag”, そしてキーワード “人工知能” という 4 つのエンティティが関わったと見なすことができる．このアクションを表現するために，我々は，このアクションそれ自体に対応する一つの *action node* を生成する．更に，このアクションに関わった 4 つのエンティティに対応する 4 つのノード，ユーザー A，URL l，キーワード “人工知能”，行為の動詞としての “tag” を生成する．ここでは，アクションに関わったエンティティを総称して *object node* と呼ぶことにする．そして，この action node と，それぞれの object node との間にエッジを張ることで，*ActionGraph* を構築する．ActionGraph は，グラフ $G = (V_{AC} \cup V_{OB}, E)$ として定義できる．ここで，V_{AC} 内の各ノードは，*action node* で表現される，ユーザーのある時点での特定のアクションに対応し，V_{OB} 内の各ノードは，*object node* で表現される，アクションに関わったエンティティに対応する．エッジ $e \in E$ は，V_{AC} 内のノードと，V_{OB} 内のノードの間にのみ張られる．すなわち，ActionGraph は，action node と，そのアクションに関わった object node の間にエッジが張られるような，二部グラフとして表現できる．図 2 は，ユーザーのアクションの例を示しており，図 1 は，対応する ActionGraph を示している．

ActionGraph の表現は，元々の行為における三つ以上の共起を保存していることが分かる．なぜなら，ある action node から張られるエッジを辿ることで，そのアクションに関わった全てのエンティティを復元することができるからである．更に，ActionGraph の表現は，各人の時間スケールを表現できていると考えられる．なぜなら，action node は，あるユーザーによって行われた特定の時間に対応しているからである．例えば，あるユーザーがある行為を今この瞬間に行い，明日また別の行為を行った場合，それらは異なる action node として表現される．各ユーザーと，特定の時間の共起を保存しており，同じユーザーでも別の時間帯に行われた行為は区別されている．

2.2 ActionGraph を用いたユーザーの興味の予測

本論文では，ユーザーがどのようなリソースに対してどの程度興味を示すかを，ユーザーとリソースの類似度を計算することで予測する．似たようなアクションを行うユーザー同士は似ているはずだ，という本論文の仮説に基づき，アクションを媒介として，ユーザーとリソースの間の類似度を計算する．具体的には，人とアイテムの関係がある時，それぞれ他方の観点から捉えて人同士/アイテム同士の類似度を測るという協調フィルタリングの基本的アイデアを，アクションとエンティティの関係にも適用する．この要求を，ActionGraph の構造を活用して実現するために，*Graph Kernel* と総称される手法を用いる．適切な Graph Kernel を用いて類似度を測ることで，二つのノード間の類似度は，そのノード間により短いパスがより多くあるほど高くなる．様々な Graph Kernel が提案されているが，その中でも，正則化ラプラシアンカーネル [Smola and Kondor2003] などの，ラプラシアン行列に基づいた Graph Kernel は，ノード間の関連度を測るのに適していると考えられる [Ito *et al.*2005]．行列 $\mathbf{M}$ の非正則化ラプラシアン行列は $\mathbf{L}(\mathbf{M}) = \mathbf{D}(\mathbf{M}) - \mathbf{M}$ と表現できる．ここで，$\mathbf{D}(\mathbf{M})$ は $\mathbf{M}$ の対角行列である．正則化ラプラシアン行列は，非正則化ラプラシアン行列の $\mathbf{L}(\mathbf{M})$ に対して，左右から $\mathbf{D}(\mathbf{M})^{-1/2}$ を掛けるることで，$\mathbf{D}(\mathbf{M})^{-1/2}\mathbf{L}(\mathbf{M})\mathbf{D}(\mathbf{M})^{-1/2}$ として得られる．正則化ラプラシアンカーネル RL_β は以下のように定義できる．

$$\mathbf{RL}_\beta(\mathbf{A}\mathbf{A}^\top) = \sum_{k=0}^{\infty}(-\beta\mathbf{L}(\mathbf{A}\mathbf{A}^\top))^k = (\mathbf{I} + \beta\mathbf{L}(\mathbf{A}\mathbf{A}^\top))^{-1} \tag{1}$$

ここで，$\mathbf{A}$ は，ActionGraph$G = (V_{AC} \cup V_{OB}, E)$ の全ノード数を n とすると，行数，列数共に n である隣接（対称）行列であり，$\mathbf{I}$ は $\mathbf{A}$ と同じ行数，列数を持つ，単位行列である．β は，拡散係数と呼ばれるパラメータである．ActionGraph に対して Graph Kernel を用いてノード間の類似度を計算するアルゴリズムは他にも考えられるが，本論文では簡単のため，正則化ラプラシアンカーネルに限定して話を進める．

3. 実験

3.1 データセット

人々の行為に着目した場合，様々なアプリケーションを利用していることが想定されるため，異なるアプリケーションからデータをアグリゲートできることが望ましい．URL は Web 上の様々なリソースを表現でき，異なるアプリケーションのデータを繋げることができるエンティティである．このため，本論文では，URL を含むアクションに着目した．Twitter から，この要求を満たす 2 つの機能，*retweet* と *favorite* をアクションデータに用いる機能として選定した．以下，簡単にこれら機能を説明する．ユーザーは，*tweet* と呼ばれる短いテキストを発信できる．tweet の中には URL も記述できるため，ここでは URL を含むような tweet のみ扱う．*retweet* は，他のユーザーの tweet を再発信する機能であり，*favorite* は他のユーザーの tweet をお気に入りとしてマークする機能である．Twitter の特定ユーザーからクロールを開始し，ユーザーの following ネットワークを 2 ホップまで辿り，得られたユーザーの，2010 年 8 月 1 日から 2010 年 8 月 30 日までのアクションデータをクロールした．データセットに用いた関係タプルを表 1 に示した．*Original-tweet-user* とは，元々の tweet を発信したユーザーのことであり，その後，他のユーザーによって，その tweet が再発信されたりお気に入りとしてマークされたりしたということである．

3.2 実験設定

ユーザーがあるリソースに対して何らかのアクションを取った場合，そのユーザーはそのリソースに対して興味を持ったと解釈できる．ここでは，ユーザーの興味を予測するというタス

表 1: 予測タスクに用いたデータセットの概要.

Action type	Facets	Tuples
Retweet	(user, URL, original-tweet-user)	14,392
Favorite	(user, URL, original-tweet-user)	25,884

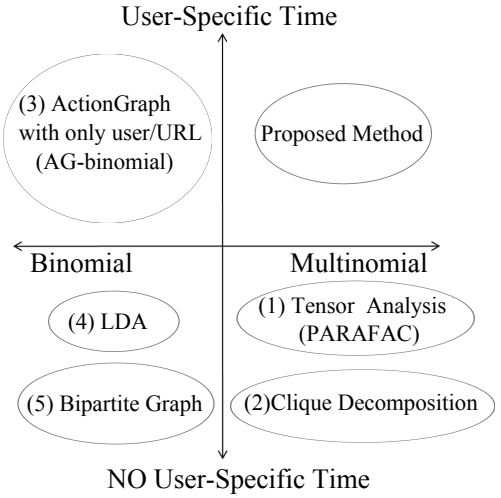

図 3: 提案手法と比較手法の関係.

クを，各ユーザーが興味を持つであろうリソースのランキングを予測する問題として定義する．入力は，表 1 に示したような関係タプルであり，この関係タプルを用いて ActionGraph を構築する．例えば，retweet アクションに対する入力のタプル（ActionGraph で object node として生成されるエンティティの種類）ならば，(user, “retweet”, URL, original tweet user となる．出力は，関係タプルに含まれていたエンティティ間の類似度である．ユーザーがあるリソースに対してどの程度興味を持つかを，ユーザーとリソース間の類似度に基づいて計算する．データの分割方法を決定するために行った予備実験で，それぞれ 3/9/27 日ごとにデータを分割して実験を行った結果，3 日ごとにデータを分割した場合の予測精度が最も良かったため，データを 3 日ごとに分割し，10 個のタイムスロットを用意するようにした．$t \in [1, 10]$ によって，タイムスロットのインデックスを示し，スロット $t \in [1, 9]$ のデータを訓練データとして用い，$t + 1 \in [2, 10]$ のデータをテストデータとして用いた．

3.3 比較手法

提案手法は，多項関係とユーザー特有の時間スケールという二つの性質をモデル化しているため，この二つの観点から比較手法を設計した．提案手法と比較手法の関係を，図 3 で示した．多項関係とユーザー特有の時間スケールのうち，どちらも扱えない手法，どちらか一方のみ扱える手法が比較手法であり，図 3 における位置によってその特性を示した．

PARAFAC はテンソル分解のスタンダードな手法である．テンソルは，多次元配列の表現であり，PARAFAC は高次元の主成分分析に相当する．PARAFAC を用いて，各次元を潜在空間に射影することができる．本実験では，ユーザーと URL は潜在空間に射影された後，それぞれのベクトルのコサイン類似度によって予測を行う．*Clique Decomposition* では，複数エンティティの共起を，グラフ上でペアワイズのエッジに分割する．例えば，あるアクション (user, “retweet”, URL) をグラフ上で表現する際には，(user, “retweet”), (user, URL), (“retweet”, URL) という三つのノードペア間にエッジが張られる．ノード間の類似度は，提案手法と同じアルゴリズムである正則化ラプラシアンカーネルによって計算する．グラフベースの三つの手法，(2),(3),(5) においては，同様にノード間の類似度は正則化ラプラシアンカーネルを用いて計算する．*AG-binomial* では，ActionGraph を構築するが，ユーザーと URL のエンティティのみを活用する．*LDA* は，*topic* と呼ばれるような，未観測の潜在変数を導入し，既知の観測データを説明しようとする，確率的生成モデルの一種である．ユーザー u があるリソース (URL) r_i に対して興味を持つ確率は，下記のように定式化できる．$P(r_i|u) = \sum_{j=1}^{Z} P(r_i|z_i = j)P(z_i = j|u)$ ここで，$P(r_i|u)$ は，あるユーザー u が与えられたときの，i 番目の リソース（URL）の条件付き確率であり，z_i は潜在トピックである．*Bipartite Graph* は，ユーザーと URL ノード間のみにエッジを張るような二部グラフである．

3.4 評価指標

精度を評価するために *R-Precision* を，データ過疎に対するロバスト性を評価するために，*Coverage* を評価指標として採用した．*R-Precision* は，正解データが R 個ある際の，ランク R での precision である．URL に対する言及はユーザーによってばらつきがあるため，R-Precision は提案タスクにおいて適切であると考えられる．*Coverage* は，テストデータにおいて，類似度を計算できた ⟨user, URL⟩ のペアのパーセンテージである．R-Precision，Coverage それぞれの値を，各ユーザー，各タイムスロットに対して平均した．

3.4.1 パラメータ調整

全データの 20% を，予備実験用のデータとして用い，R-Precision の観点から，各手法についてパラメータを調整した．提案手法においては，式 1 におけるパラメータ β を $0.01, 0.05, 0.1$ と変化させた．結果はほぼ変わらなかったが，$\beta = 0.1$ と 0.01 の際の結果が同じであり，若干良いものだったため，β を 0.1 に設定した．他のグラフベースの手法はこのパラメータ設定に従う．正則化ラプラシアンカーネルは，パラメータ β に対して比較的安定していることが指摘されている [Ito *et al.*2005] が，実際，今回の実験でもパラメータ β に対する安定性が見られた．PARAFAC では潜在空間の次元を $200, 400, 600, 800$ と変化させ，最も良かった 600 を採用した．LDA では，トピック数を $100, 200, 400, 600$ に変化させた結果 400 を得た．LDA では，他に α と β と呼ばれるパラメータが存在する．まず β を $0.01, 0.05, 0.1$ と変化させ，0.01 を得た．α をトピック数に応じて変化させることが有効であることが指摘されている [Wallach *et al.*2010] ため，本実験でも，利用した pLDA というライブラリ[*1]の実装に従い，各トピックに対して α を $\frac{50}{topicNumber}$ のように変化させた．

3.5 結果と分析

結果を表 2 に示した．R-Precision の観点からは，ActionGraph は他の比較手法より，概ね良い値になっていると言える．この理由としては，多項関係とユーザー固有の時間スケールの両方が，ユーザーの興味を捉えるのに有用であることが考えられる．Coverage に着目すると，ActionGraph は，テンソル分解の標準的な手法である PARAFAC と，二項関係を扱う二つのグラフベースの手法，Bipartite Graph と AG-binomial に対して，有意に良い値となっている．Actiongraph は，可能なほぼ全てのユーザー/URL のペアに対して類似度を計算できていることが分かる．ユーザーや URL が過疎である状況でも，

∗1 www.code.google.com/p/plda

表 2: 各手法における，予測性能の平均.

	Proposed Method	PARAFAC	Clique Decomposition	AG-binomial	LDA	Bipartite Graph
R-Precision	7.6±3.3 %	3.4±2.1 %	4.2±1.6 %	4.2±2.6 %	4.3±1.4 %	2.0±1.1 %
Coverage	99.8±0.0 %	43.6±6.7 %	99.8±0.0 %	41.9±7.9 %	99.0±0.2 %	56.4±10.8 %

アクションの種類（retweet や favorite）は多くの action node 間で共有されるので，アクションの種類が，あらゆるノード間の橋渡しとなり，データ過疎を緩和していると考えられる．ゆえに，ActionGraph はデータ過疎に対してロバストであると期待できる．

4. 関連研究

4.1 多項関係の分析

多項関係の分析により，三つ以上のエンティティの共起を捉えることが可能になる．これは，様々なケースにおいて有効であると期待される．例えば，異なるユーザーが同じドキュメントに対して同じキーワードでタグ付けをしたとしても，そのタグのキーワードの意味はそれぞれのユーザーによって異なっていると考えられるためである．多項関係を扱う既存手法としては，ペアワイズに分割する手法やテンソルによるモデル化が挙げられる．ペアワイズカーネルの [Ben-Hur and Noble2005] は，タンパク質間の相互作用を予測するために提案されたもので，同様のカーネルが，エンティティの名前解決タスクや，協調フィルタリングタスクにおいて提案されている．[Zhou *et al.*2008] は，著者と論文，会場という三つのエンティティを扱い，それらを三つの二部グラフに分割している．しかし，これらのペアワイズの手法では，元々生じていた三つ以上の共起の情報が欠落してしまう．一方，テンソルベースの手法も研究が盛んに行われている．[Lin *et al.*2009] は NMF (Non-negative matrix factorization) [Lee and Seung2001] をテンソルの場合にも一般化し，social media でのコミュニティの変遷を追う手法を提案している．これらペアワイズの手法やテンソルベースの手法は有望だが，ユーザー固有の時間スケールを明示的にモデル化していない．

4.2 時間進化するデータの分析

時間をグローバルな次元として扱う手法は多々あった．しかし，冒頭で述べたように，時間をよりローカルな，各ユーザー固有のものとして扱う方が好ましい場合が存在する．[Xiang *et al.*2010] は，この問題に対する新しいアプローチであり，同じユーザーに近い時間帯で購入されたアイテムを繋げる *session node* を導入している．しかし，このアプローチでは多項関係を扱えていない．

5. むすび

本論文では，多項関係とユーザー固有の時間という，二つの性質を明示的にモデル化するグラフ表現，ActionGraph を提案した．提案手法は，既存の手法と比較して予測精度とデータ過疎に対するロバスト性の観点から，高い性能を示した．future work としては，ユーザー固有の時間をより明示的に活用することが考えられる．例えば，あるユーザーによって，近い時間帯に行われた action node を一つの action node として表現することで，ユーザーの短期的な嗜好をより適切に捉えることができると期待できる．

参考文献

[Bader and Kolda2006] Brett W. Bader and Tamara G. Kolda. Algorithm 862: Matlab tensor classes for fast algorithm prototyping. In *TOMS 32(4)*, pages 635–653, 2006.

[Ben-Hur and Noble2005] Asa Ben-Hur and William Stafford Noble. Kernel methods for predicting protein-protein interactions. In *Proc. 13th International Conference on Intelligent Systems for Molecular Biology. Bioinformatics 21*, pages i38–i46, 2005.

[Blei *et al.*2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Ito *et al.*2005] Takahiko Ito, Masashi Shimbo, Taku Kudo, and Yuji Matsumoto. Application of kernels to link analysis. In *Proc. 11th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2005)*, pages 586–592, 2005.

[Lee and Seung2001] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances Neural Information ProcessingSystems*, 13:556–562, 2001.

[Lin *et al.*2009] Yu-Ru Lin, Jimeng Sun, Paul Castro, Ravi Konuru, Hari Sundaram, and Aisling Kelliher. Metafac: community discovery via relational hypergraph factorization. In *Proc. 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2009)*, 2009.

[Smola and Kondor2003] Alex J. Smola and Risi Kondor. *Kernels and regularization on graphs.* Springer, 2003.

[Wallach *et al.*2010] Hanna M. Wallach, David Mimno, and Andrew Mccallum. Rethinking lda: Why priors matter. In *In Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS 2009)*, 2010.

[Xiang *et al.*2010] Liang Xiang, Quan Yuan, Shiwan Zhao, Li Chen, Xiatian Zhang, Qing Yang, and Jimeng Sun. Temporal recommendation on graphs via long- and short-term preference fusion. In *Proc. 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2010)*, 2010.

[Zhou *et al.*2008] Ding Zhou, Shenghuo Zhu, Kai Yu, Xiaodan Song, Belle L. Tseng, Hongyuan Zha, and C. Lee Giles. Learning multiple graphs for document recommendations. In *Proc. 17th International World Wide Web Conference (WWW 2008)*, pages 141–150, 2008.