ARG WI2 No.14, 2019
WI2-2019-13

Hawkes過程とトピックモデルによる検索数の予測

岩田　晟[†,a]　　江口　浩二[††,b]　　藤田　澄男[†††,c]

† 神戸大学大学院 システム情報学研究科　†† 広島大学 情報科学部　††† ヤフー株式会社

a) *191xe03x@stu.kobe-u.ac.jp*　b) *eguchi@acm.org*　c) *sufujita@yahoo-corp.jp*

概要

インターネット上で検索クエリを投入するとき，あるクエリの影響によって他のクエリの発生が起きることがある．本稿ではそのような検索クエリの影響を表すため，点過程の一種 である Hawkes 過程の枝因子に着目し，その枝因子を用いて，どれほど今後，検索されかを予測する．Hawkes 過程の枝因子はある事象の発生によりどれほどの事象数が発生するかを表す．LDA-Hawkes モデルにて各クエリがどのトピックに属するか推定したのち，枝因子を用いることで検索クエリの今度発生する事象数，すなわち，今後，どれほど検索されるかを予測する．このとき予測され検索数はトピックごとの検索数であるため，分野 (トピック) ごとの流行を予測することもできる．本稿では，Yahoo!Japan のクエリログの一部を用いた実験の結果を報告する．

キーワード　トピックモデル, Hawkes 過程 , 検索数

1　はじめに

スマートフォンやソーシャルメディアの発展とともに，インターネット上で情報検索が行われる．それに伴い，どの分野でどの単語がこれからどれほど，検索されるかを予測することでその単語の表す事物のトレンドを掴むことができると考えられる．単語の流行を知ることで，適切なマーケティングやビジネスの展開に繋げることが期待される．

そのために，検索クエリがどのトピック (非明示的なカテゴリ) に属しているかを判別し，各単語の検索数を予測することを本稿では目標とする．そのような潜在的なトピックを推測する手段としてトピックモデルの一種である Latent Dirichlet Allocation(LDA)[1] がある．検索クエリに LDA を活用する手段として，ユーザの検索タスクが時間的に変化することを仮定し，固定の時間長で分割されたデータから学習する LDA がある．しかし，固定時間長でデータを区切ると，異なるデータ群のクエリの共起が無視される. そのため，各クエリの共起性を表す手段として，ある事象が他の事象の発生確率を上げるような影響を表す Hawkes 過程がある．LDA と Hawkes 過程を組み合わせた LDA-Hawkes モデル [2] がある．また，Hawkes 過程の枝因子を用いることである事象によって直接発生する事象数を予測することができる [3].

本稿では LDA-Hawkes モデルを用いてクエリの潜在トピックを推定し，その後，Hawkes 過程の枝因子を用いて検索数を予想する.

2　関連研究

地震やクエリなどを表す統計的手段として点過程がある．最も単純な点過程として Poisson 過程がある．Poisson 過程はそれぞれの事象が独立に発生すると仮定しているため，事象の発生に過去の事象の影響を考慮しない．しかし，実際に検索を行うにあたり，過去の検索クエリの影響を受けて検索事象が発生することがある．例えば,「大阪 観光地」というクエリで検索して，”通天閣” について知り，その後,「大阪 観光地 通天閣」というクエリで検索する場合，後のクエリは前のクエリに影響されたと考えられる．そうした過去からの影響を考慮する Hawkes 過程について，Poisson 過程と共に以下で説明する．その後，従来の研究における取り組みに触れる．

2.1　Hawkes 過程

$N(t)$ は時間 t 以前に起きた事象数をとし，$h > 0$ とする．このとき，時間 $t \sim t + h$ の間に事象が起きる確率は確率強度関数 λ を用いて次のように表される．

$$P(N[t+h] - N[t] = 1) = \lambda(t)h \tag{1}$$

この λ によって特徴付けられる Poisson 過程を次のように過去の事象から影響を受け，自己的もしくは相互的に励起するように拡張したのが Hawkes 過程 [4] である．

$$\lambda(t) = \mu(t) + \sum_{i:t>T_i} \Phi(t - T_i) \tag{2}$$

$$\Phi(t) = \beta\kappa(t) \tag{3}$$

ここで μ は基本強度を表し，T_i は時刻 t より以前に発生した i 番目の事象時刻，β は影響の大きさ，κ 関数は影響の時間減衰を表している．また，過去の事象の影響力を表すメモリカーネル関数 Φ を含む第二項は過去の事象らの影響の総和である．

2.2 Hawkes 過程の枝因子を用いた Twiiter の拡散予想

Rizoiu ら [3] は Hawkes 過程の枝因子を用いて Twitter のリツイート (他のユーザにより再投稿) を予測している．枝因子とはある事象の発生により直接的に発生する平均予測事象数である．枝因子 n^* は式 (2) の Hawkes 過程のカーネル関数 Φ から以下のように求めることができる．

$$n^* = \int_0^\infty \Phi(\tau) d\tau \tag{4}$$

この枝因子から一つ，次の世代の事象数がわかる．例えば，i 世代目の事象数 A_i は次の様に求まる．ただし，$A_0 = 1$ とする．

$$A_i = A_{i-1} n^* = A_{i-2}(n^*)^2 = ... = A_0(n^*)^i = (n^*)^i, i \geq 1 \tag{5}$$

そして，ある事象の発生によりその後，発生する全事象数 N_∞ は次の様になる．

$$N_\infty = \sum_{i=0}^{\infty} A_i = \frac{1}{1-n^*}, n^* < 1 \tag{6}$$

ここで,，$n^* < 1$ である必要がある．なぜなら，n^* とは次に起きる平均事象割合であり，$n^* > 1$ のときは情報が収束せず，拡散され続けることを意味するからである．

Rizoiu ら [3] では N_∞ の予測を用いることで Twitter の拡散具合を予測している．その中ではカーネル関数が精度に影響することが述べられており，べき則カーネルの方が指数カーネルよりも Twitter データでは精度が高いことが示されている． 本研究では検索クエリデータを対象にすることで LDA によってトピックを求めたあと，クエリの検索数に対して行うことで，トピック内検索ワードの流行を予測することが期待できる．

2.3 検索クエリデータに対するアプローチ

Li ら [2] は検索クエリデータに対して LDA と Hawkes 過程を組み合わせたモデルを提案した．彼らが検索タスクの識別を意味的類似度に着眼して評価しているのに対し，本稿ではトピック毎の検索クエリ数の予測が目的であり，そのために LDA-Hawkes モデルを用いる．

3 検索クエリデータを用いたトピック毎の検索数予測

本稿では，検索データを対象に，LDA-Hawkes モデルによる各クエリがどのトピックに属しているかを推定し，その後，今後発生する事象数を Hawkes 過程の枝因子に よって予測する．本稿の目的は，トピック内でのクエリ予測を行うことである．それにより，トピック (非明示的なカテゴリ) ごとの流行り廃りを掴むだけでなく，ある単語が異なるトピックに変化するようなことも観測できると期待される．

3.1 検索データに対しての枝因子と事象数の予測

LDA-Haekes モデル [2] を検索データに対して学習し，得られたパラメータを基にあるクエリの発生により発生する事象数を予測するために，ユーザのクエリ毎の各トピックの確率強度を求める．ユーザ m の文書 n のトピック k の確率強度は次のようになる．

$$\lambda_{m,n,k}(t_{m,n}) = \sum_{t_{m,l}<t} \phi_{m,n,k} * \phi_{m,l,k} * \beta_m * \kappa(t_{m,n} - t_{m,l}) \tag{7}$$

$\phi_{m,n,k}$ は LDA-Hawkes によって学習されたユーザ m の n 番目のクエリの k 番目のトピックに割り当てられる確率である．ここで $\phi_{m,n,k} * \phi_{m,n,l}$ はカーネル関数の大きさ一部とも考えることができるため，式 (3) のカーネルの大きさ β を今回は以下の様に平均値をとりユーザ毎に固定値とする．

$$\beta = \beta_m \frac{\sum_l^{n-1} \phi_{m,n,k} * \phi_{m,l,k}}{n-1} \tag{8}$$

このように設定すると Hawkes 過程の枝因子は次の様になる．

$$n^*_{m,n,k} = \int_0^\infty \beta\kappa(\tau) d\tau \tag{9}$$

これにより，k 番目のトピックに属しているユーザ m 番目の n 番目のクエリの今後発生する事象数 $N_{m,n,k}$ が次の様に求められる．

$$N_{m,n,k} = \sum_{i=0}^{\infty} A_i = \frac{1}{1-n^*_{m,n,k}}, \qquad n^* < 1 \tag{10}$$

特定の単語の未来の事象数を求めるために，各ユーザの特定の単語が含まれる最後のクエリの $N_{m,n,k}$ を求め，その総和を特定の単語が今後発生する事象数とする

3.2 LDA-Hawkes のカーネル関数

式 (4) の時間減衰関数 κ に対して，Hawkes 過程では一般的な指数カーネル [4] と地学 [5] やソーシャルメディア [6] の分野で使われるべき則カーネルを用いる．

$$\kappa(\tau) = \mathrm{e}^{-a\tau} \tag{11}$$

$$\kappa(\tau) = (\tau + c)^{-(\theta+1)} \tag{12}$$

本実験ではこれらのカーネル関数のパラメータ (指数カーネル式 (11) の a，べき則カーネル式 (12) の τ と θ) に対して，いくつかの固定値を用いてそれらと比較する．

表 1 各カーネル関数と 15 時以降の検索とのケンドールの順位相関係数

data	a=1.0	a=0.5	a=0.1	a=0.05	a= 0.01	べき則
10 人	-0.059	-0.015	-0.044	0.029	-0.044	0.0147
20 人	0.045	0.267	0.176	0.202	0.176	0.281
100 人	0.404	0.404	0.404	0.404	0.404	0.404

表 2 各カーネル関数とランキングデータとのケンドールの順位相関係数

data	a=1.0	a=0.5	a=0.1	a=0.05	a= 0.01	べき則
10 人	0.098	0.098	0.098	0.142	0.120	0.164
20 人	0.309	0.263	0.263	0.216	0.239	0.368
100 人	0.415	0.461	0.461	0.461	0.461	0.461

4 実験

4.1 データセット

データセットには Yahoo! Japan の検索クエリデータを用いる．検索クエリデータは，暗号化された Yahoo! ID，unix timestamp，クエリ，Yahoo! Japan の提供するサービスのドメインで構成されている．今回はドメインを”search.yahoo.co.jp”のみに限定し，日にちを 2016 年 12 月 25 日に設定した．全検索データから無作為にユーザを抽出した場合，少人数では入力データに偏りが生じたため，本実験では,「有馬記念」と検索した人のうち，50 前後の件数のクエリを出している人物を 10 名分抽出したデータと，20 名分抽出したデータ,100 名分抽出したデータの 3 セットを用いた．ここで「有馬記念」としたのは．2016 年 12 月 25 日には競馬の「有馬記念レース」があり，競馬のレースが終われば情報検索数が収束し始めると考えたからである．正解データを二つ設定した．一つ目は，一番大きなレースが 15 時ごろに始まるため，検索データセットのうち，14 時までのデータを学習データとし，それ以降の検索データを正解データとする．二つ目は，2016 年 12 月 25 日の Yahoo!Japan 社内サービスのランキング検索ランキングデータと比較を行う．

入力データの前処理として，MeCab で形態素解析を行い，その後，以下のような処理を行った．

1. ”助詞”などを除いて”名詞”，”動詞”，”形容詞”だけを抜き出す
2. URL のクエリを除く
3. 英単語を全て小文字に変える
4. 日本語，英語のストップワードの単語を取り除く

4.2 実験設定

LDA-Hawkes モデルでは上記のデータセットに対して変分ベイズ法により，未知パラメータを推定した．これにより得られたパラメータを用いることで特定の単語を含むクエリの将来の事象数を予測する．ただし，LDA-Hawkes モデルの LDA 由来の超パラメータの値のうち，トピック分布のディリクレ事前分布の超パラメータを $\alpha = 0.1$，単語分布のディリクレ事前分布の超パラメータを $\alpha' = 0.01$ とした．また，Hawkes 過程のカーネル関数のパラメータは式 (11) の指数カーネルの場合，$a = 1$, 0.5, 0.1, 0.05, 0.01 と 5 つの値で，べき則カーネルの場合，$c = 58$,$\theta = 0.64$ として比較した．評価方法としては，正解データと予測された検索数に対してケンドールの順位相関係数を用いて比較した．

4.2.1 ケンドールの順位相関係数

ケンドールの順位相関係数はデータ点数を N とする 2 つのデータ集合毎に数値によるランキング ($1 \sim N$ 番) をつけ，その順位によって算出する．以下はケンドールの順位相関係数 $\tau_{x,\,y}$ の定義である．

$$\tau_{x,\,y} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{i-1} D(x_i,\,x_j)D(y_i,\,y_j)}{{}_N\mathrm{C}_2} \tag{13}$$

$$D(a,\,b) = \begin{cases} 1 & (a > b) \\ -1 & (a < b) \end{cases} \tag{14}$$

ここで関数 D はデータの方向性を表す関数である．つまり，データ $X = (x_1,...,\,x_i,...,\,x_N)$ とデータ $Y = (y_1,...,\,y_i,...,\,y_N)$ の各インデックスにおける値の大小の順位の方向が同じ場合は 1 を返し，逆の場合は-1 を返す．また，ケンドールの順位相関係数 τ はピアソンの相関係数と同様に $-1 \leq \tau \leq 1$ で表され，二つのランキングが独立しているほど 0 に近く，相関が高いほど，± 1 に近い値を示す．

5 結果

三つのデータセットに対して二つの正解データを基準とし，指数カーネルが $a = 1, 0.5, 0.1, 0.05, 0.01$ の場合

とべき則カーネルが $c = 58, \theta = 0.64$ の場合の実験の結果を表 1，表 2 に記す．

表 1, 表 2 ともに人数の多いデータセットの方が精度が良い．これは LDA-Hawkes モデルが事象間の影響を考慮することで単語の共起性をユーザ間をまたいで学習することに起因していると考えられる．また，べき則カーネルが概ねどのデータセットに対してよい．Twitter の拡散を予測する研究 [3] でも指数カーネルよりべき則カーネルの方がよく．これと共通した傾向が認められる．今回のデータセットに対しては指数カーネルよりベキ則カーネルの方がうまく影響の減衰が表せていることがわかる．また，データセット人数を 100 人にした場合，全体的に精度は上がったが，ケンドールの順位相関係数の値が近しくなり，カーネル関数の優劣が見られなかった．より詳細に予測検索数自体で比較できる別の指標で評価したいと考えている．

6 おわりに

本稿では検索クエリデータを LDA-Hawkes モデルを用いて分類した後，Hawkes 過程の枝因子を用いて，検索クエリの検索数を予測した．評価手法として，ケンドールの順位相関係数を用いて，指数カーネルとべき則カーネルを比較し，パラメータにより有効性が大きく変化することがわかった．今回は Hawkes 過程のカーネル関数のパラメータを経験則的に置いたが，それらのパラメータの最尤推定法 [5] を行うことでパラメータを推定することを考えている．また，検索クエリの流行は生き物のように常に変動しており，緩やかなときもあ れば，突発的に変化するときもある．緩やかな場合は，今回のようなバッチ法でも問題ないが，突発的に変化する場合は学習時間に時間がかかり，対処している間に流行 が変わる可能性がある．そのため，文献 [7] でのオンライン学習を LDA-Hawkes モデルに適応することで逐次的な学習が可能になり，そのような突発的な変化に対応することが考えられる．

謝辞

本研究において，Yahoo! Japan 研究所の検索クエリデータを利用させていただきました．深く感謝いたします．

参考文献

[1] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent dirichlet allocation, Proc. of Journal of Machine Learning Research, 3,pp. 993-1022，2003.

[2] Li, L., Dong, A., Chang , Y. and Zha ,H.: Identifying and labeling search tasks via Query-based Hawkes process, Proc. of KDD,pp.731-74-, 2014.

[3] Rizoiu, MA., Lee, Y. Mishra, Y., and Xie, L.: A tutorial on Hawkes process for events in social media, arXiv:stat.ML ,1708.06401, 2017.

[4] Hawkes, A. G.: Spectra of some self-exciting and mutually exciting point processes. Proc. of Biometrika. Vol. 58, pp. 83–90, 1971.

[5] Ozaki, T.:Maximum likelihood estimation of Hawkes’ self-exciting point processes. Annals of the Institute of Statistical Mathematics, Vol.31, pp.145 – 155, 1979.

[6] : Rizoiu, MA., Xie, L., Sanner, S., Cerbrian, M., Yu, H. and Henteryck, P. V.:Hawkes Intensity Processes for Social Media Popularity. International World Wide Web Conference Committee(IW3C2), PP735-744,2017.

[7] Hoffman , M.D., Blei, D. M., Bach, F. :Online Learning for Latent Dirichlet Allocation. Proc. of NIPS,2010.