1L3-5

消費に関する検索共起語から市場を観測する手法に関する研究

Observe Market with Co-occurrence of Consumption-Related Word in Search Queries

大野 峻典[*1] 上野山 勝也[*1] 松尾 豊[*1]
Shunsuke Ono Katsuya Uenoyama Yutaka Matsuo

[*1] 東京大学
The University of Tokyo

Recent studies have shown that you can predict or observe a market with people's search logs on search engines. They analyze the volume of search queries related to specific keywords. Search logs include variety of queries with different intents, so you need to extract relevant queries from them and choose suitable prediction models for analysis. In this study, assuming co-occurrence words in search queries imply the intents of those queries, we used different combinations of search queries for prediction and compared the predictions' accuracy. In addition, we applied 4 types of models with those combinations of search queries and compared the predictions' accuracies. In conclusion, we implied the possibility of improvement of prediction accuracy by extracting logs and choosing models based on co-occurrence words.

1. はじめに

近年，ウェブ上における人々の大量の行動データに対し，統計処理を加え，実世界を観測できることが言われてきた.[Sakaki 2012] 特に検索クエリデータを用いた研究では，不動産の売上や失業率など，経済現象・社会現象がリアルタイムに計測できることが示されてきた.[Choi 2009a, Choi 2009b, Goel 2010]

こうした研究では検索ログに対して適したロジックを基に処理を行うことで，対象の事象を観測する．検索ログに用いるロジックとして，全検索クエリログから観測モデルの素性に用いる検索クエリの抽出と，観測モデルの選択，の二点を適切に行う必要性が認識されている．しかし，先行研究では，ヒューリスティックな素性抽出とモデル選択が行われており，それぞれにフォーカスした考察が加えられたものは無く，知見が蓄積されていない．そこで本研究では，この二点に関して有用な考察を加えることを目的とした．特に検索クエリの含む共起語がそのクエリの検索行動の意図を示唆すると考え，共起語に着目して，素性抽出とモデル選択に関する考察を行った．

検索クエリ群からの素性抽出に関しては，共起語から推測される検索行動のタイプにより検索クエリを分類し，それぞれのクエリを用いた場合の観測精度を比較した．また，モデル選択に関して，各素性に対して異なる観測モデルを適用し観測精度を比較した．結果，特定の検索行動によって生じる検索クエリが観測に適していること，また，各クエリに対して適したモデルが存在することがわかった．それぞれ背景にある理由を考察し，検索クエリを用いた事象の観測のための共起語を用いた素性抽出とモデル選択における有用な知見を得た．

具体的には，本研究における実験では，大手通信事業会社である「NTTドコモ」「au by kddi」「ソフトバンク」の三社の携帯端末契約数とその増減数を対象とし，「Yahoo! Japan」の検索ログを用いて観測を行った．素性抽出に関して，クエリを検索共起語から推測できるユーザの契約状態により分類し，各状態におけるクエリを用いた場合の，契約数と契約増減数の観測精度を比較した．結果，契約数観測においては，携帯端末を既に契約しているユーザによる検索クエリを用いて観測を行うことで比較的観測精度が高くなることがわかった．契約の増減数観測においては，携帯端末の契約を検討している状態のユーザによるクエリを用いると観測精度が高くなることがわかった．また，契約数観測におけるモデル選択に関して，各社の携帯の機能に関する検索クエリのように検索数が契約中ユーザ数の増加に比例して増えるクエリを素性に用いる場合，線形のモデルが観測に適すことがわかった．それに対して，解約に関するクエリなど，契約中ユーザ数の伸びに対して対数関数的に検索数が増加しているクエリを用いる場合は，非線形モデルが観測に適していることがわかった．

2. 関連研究

「Yahoo!」の検索ログデータを用いて，映画の興行収入や音楽の人気の予測を行う研究[Goel 2010]では，観測に用いるクエリの絞込を，検索エンジンによる検索結果や，「Yahoo!」の特定のサービスにおける検索クエリに限定することで行っているが，これらはそうした追加データを必要としたヒューリスティックな方法であり，また，クエリの抽出方法に関する考察はされていない．検索ログデータからインフルエンザ流行の観測を行う研究[Polgreen 2008]では，インフルエンザに関わる単語を含む検索クエリの中から，共起語に着目してヒューリスティックに不適切なクエリを排除しており，共起語に着目したクエリの分類・抽出の有用性に関する考察はしていない．

適した観測モデルの選択に関して，検索ログを用いた市場観測を行っている従来の研究[Choi 2009a, Choi 2009b, Goel 2010]では，観測対象に応じて線形，非線形のモデルをヒューリスティックに使い分けているが，モデル選択にフォーカスした考察はしていない．

3. 提案手法

観測対象の指標を目的変数とし，あるクエリの検索回数を説明変数として，複数のモデルを用いた観測を行う．

3.1 データセット

本手法では，日本最大級の検索エンジン「Yahoo! Japan」における検索ログの全量データ [1] より，日毎に検索されたクエリとそれぞれの検索回数の集計データを用いる．日毎に存在するクエリの種類は数億オーダーである．

3.2 観測対象に関連する検索クエリの獲得

観測対象の呼称を含む検索クエリを，その観測対象に関連す

連絡先：大野 峻典，東京大学，
shunsuke.ono@weblab.t.u-tokyo.ac.jp

[1] 本研究は Yahoo! Japan を運営するヤフー株式会社との研究プロジェクトの一貫で行われている．

るクエリと考え，取得する．例えば，「NTT ドコモ」の携帯端末契約数を観測する場合は，“docomo”や”ドコモ”という呼称を含む「ドコモ wifi」や「docomo メール」といったクエリを検索クエリログとして取得する．

3.3 各クエリを素性とした観測モデルの精度比較

前項で獲得した全検索クエリログから観測に適したクエリ抽出に関する知見を得るため，各クエリを観測モデルの素性とした場合の観測精度を比較する．

共起語と精度の関連性に着目するために，まずクエリを共起語から推測される検索行動の意図に応じて分類し，それぞれのクエリを素性とし線形モデルを用いて予測を行った場合の予測精度を比較する．

線形回帰モデルにおけるパラメータをθ_0, θ_1として検索量 X を素性として契約数を予測する仮説関数$h_\theta(X)$は以下のように表され，最小二乗誤差を最小化するようにパラメータを学習する．

$$h_\theta(X) = \theta_0 + \theta_1 X$$

また，素性に用いるクエリを複数個にした場合の観測精度をクエリの個数毎に比較する．複数クエリの検索量をそれぞれX_1, X_2, X_3 ...とした場合，仮説関数は以下のように表され，最小二乗誤差を最小化するようにパラメータを学習する．

$$h_\theta(X) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \cdots$$

3.4 各モデルを用いた精度比較

各クエリによる素性に対して，異なるモデルを用いた場合の精度比較を行った．具体的には，それぞれの素性に対して，線形回帰，線形サポートベクトル回帰，RBF カーネルを用いたサポートベクトル回帰，ニューラルネットワークを用いて観測を行う．

4. 実験

4.1 観測対象の設定

観測の対象は携帯電話事業者ドコモ，au，ソフトバンクの 3 社の携帯電話契約数と，契約増減数とした．データは一般社団法人 電気通信事業者協会 [2] によって公開されているものを用いた．携帯事業者に関する検索には，端末契約をしていることを示唆する検索に加え，様々な意図の検索が存在しうるため，本手法により共起語による意図の分類と抽出を検討する対象としてふさわしいと考えた．また，国内における検索ボリュームが十分に存在し，時系列で十分な量の正解データが公開されていることも選定理由である．

2006 年 3 月から 2014 年 3 月における月次の累計契約数データを用いた．目的変数としては，各社の「月次の累計契約数」「月次の累計契約の増減数」の 2 つを設定した．

Yahoo! Japan の検索エンジンにおける日毎の検索クエリと検索回数データも同時期のものを用いた．ただし，計算量の都合上，1/100 程度のデータ量になるように，ランダムにサンプリングを行った．

4.2 各クエリを素性とした観測精度結果

通信事業者に関する検索クエリを，ユーザの契約状態の段階で分類して考えた．「興味」段階，「契約」検討段階，「利用」段階，「解約」検討段階，契約状態と「無関係」なクエリの 5 つに分類した．ドコモに関する結果を表 1 に示した．

表 1. クエリの契約状態に応じた分類

興味	契約検討	利用	解約	無関係
docomo cm, ドコモ cm	ドコモ 新機種, ドコモ 携帯 価格, ドコモショップ, ドコモ 携帯ランキング	マイドコモ, mydocomo, ドコモ 料金, domcomo id, ドコモ 着信拒否	ドコモ 解約, ドコモ mnp	ドコモ 株価, ntt ドコモ 株価, ドコモ 新卒

ドコモ契約数の観測精度でクエリを分類したものを一部表 2 にまとめた．同様に契約数の増減に関して表 3 にまとめた．

表 2. 契約数の観測精度の高さによりクエリを 10 グループに分類（Group0 が最も精度が高い）

Group0	Group1	…	Group8	Group9
ドコモ 着信拒否 mydocomo マイドコモ ドコモ 料金 docomo id	ntt ドコモ 料金 ドコモ スマートフォン docomo カード ドコモケータイ ドコモ東海	…	ドコモ 料金プラン ドコモ ポイント ドコモ 携帯 機種 ドコモ	ドコモ 株価 ドコモ cm ドコモショップ 営業時間 ntt ドコモ 株価

表 3. 契約数増減の観測精度の高さによりクエリを 10 グループに分類（Group0 が最も精度が高い）

Group0	Group1	…	Group8	Group9
ドコモショップ ドコモ cm ドコモ 携帯ランキング ドコモ 機種変更	ntt ドコモ 料金 ドコモ スマートフォン docomo カード ドコモプレミアムクラブ	…	ドコモ 料金プラン ドコモ ポイント ドコモ 携帯 機種 ntt ドコモ 株価	ドコモ 株価 ドコモ オンラインショップ docomo id ドコモ 解約

表 1, 表 2 より，ドコモの契約数予測において精度の高いクエリ，低いクエリのユーザの契約状態をみると，マイページや機能・情報確認等「利用」段階のクエリが精度の上位に来ており，株価等契約状態と「無関係」なクエリが下位に来ていることがわかる．対して，表 1, 表 3 より，契約数の増減観測では，「契約」検討段階のクエリの精度が高く，契約状態と「無関係」なクエリは精度が低くなっていることがわかる．

また，素性に用いるクエリを複数個にした場合のドコモの契約数の観測精度に関して，クエリの個数，各クエリ個数において最高精度の素性を構成するクエリの組み合わせ，精度の増減を一部表 4 にまとめた．

表 4. 複数次元の素性による観測精度

個数	最高精度の素性を構成するクエリ	精度
4	my docomo, docomo id, ntt ドコモ 料金, docomo 迷惑メール	増
5	my docomo, docomo id, my ドコモ, ntt ドコモ 料金, docomo 迷惑メール	一定
6	ドコモ 着信拒否, my docomo, docomo id, ntt ドコモ 料金, my ドコモ, docomo 迷惑メール	減

[2] http://www.tca.or.jp/

クエリの個数が 5 個に達するまで精度は上がり，それ以降精度は下がった．

4.3 各モデルを用いた観測精度結果

クエリ毎に，観測に適したモデルは異なった．線形回帰モデルが適したケースを図1に，非線形回帰モデルが適したケースを図 2 に示した．プロットは，各月の検索数と契約数を表し，直線，曲線は予測値である．

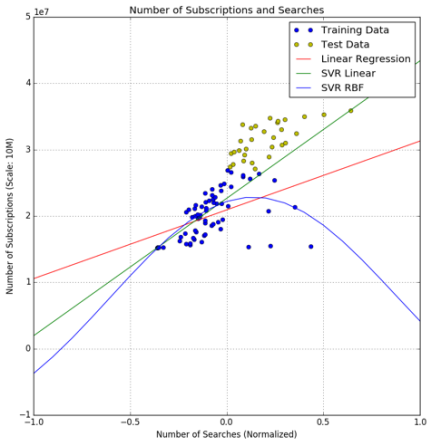

図 1「ソフトバンク メール」検索回数(横軸)と契約数(縦軸)

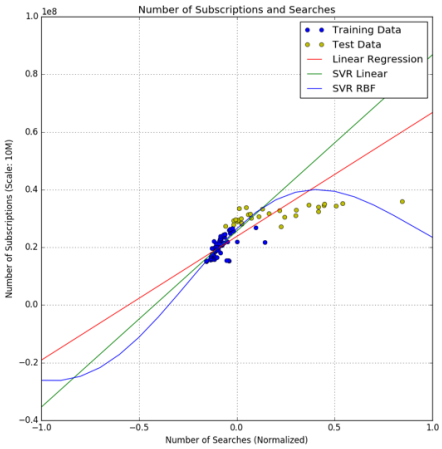

図 2「ソフトバンク 解約」検索回数(横軸)と契約数(縦軸)

ニューラルネットワークによりドコモ契約数を予測した場合の，正解値と予測値の時系列推移を図 3 に示した．プロットが各月の契約数であり，曲線は予測値である．なお，プロットの前半70%が訓練データである．

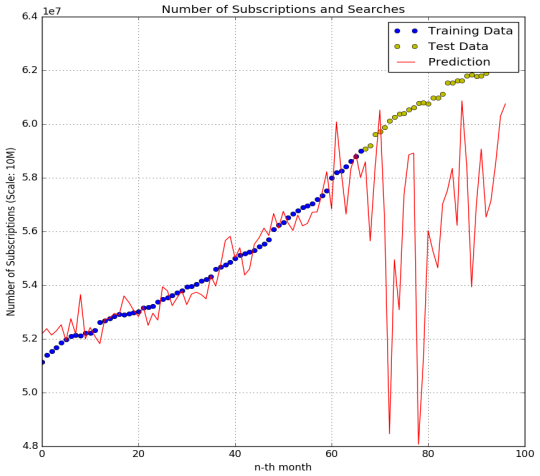

図 3. ドコモ契約数(縦軸)とニューラルネットワークによる予測値の時系列(横軸)の推移

5. 考察

契約状態で分類したクエリとそれぞれの精度から，観測の素性に用いるクエリ抽出において，契約数の観測では「利用」中のクエリ，契約増減の観測では「契約」検討中のクエリ，というように観測に適した意図を示す共起語を含むクエリを抽出すべきだと示唆できる．

表 4 より，複数のクエリを素性に用いる場合のクエリの組合せに関しては，精度が増加している 4 次元まではクエリの意味に重複は無いが，6 次元以降「my docomo」「my ドコモ」とクエリの意味に重複がうまれ，精度が減少していることがわかる．これより，クエリの意味に重複がない範囲で，適したクエリを組み合わせることで精度が上がると示唆できる．

図 1 のようにモデル選択では，クエリによって精度の高いモデルが異なることがわかる．線形モデルが適した場合，「ソフトバンク メール」のように契約数の増加に応じて線形に検索数が伸びるクエリであることがわかった．また，ノイズデータを含むケースに，線形では過学習を起こしにくく，比較的よく観測できることがわかる．非線形モデルが適した場合は，「ソフトバンク 解約」のように，ユーザの増加に対して，対数関数的に検索回数が変化するクエリであるとわかった．これより，観測対象と素性に用いるクエリの検索行動の関係性をよく説明したモデル選択を共起語に着目して行うことで，精度が改善できる可能性が示唆できる．

ニューラルネットワークのように，モデルのキャパシティが高くなるにつれ，観測精度が下がる傾向にあった．図 2 より，ニューラルネットを用いた場合，訓練データにはよくフィットしているが，テストデータで観測がうまくいっていないことがわかる．これは，観測対象に対して，モデルのキャパシティが高すぎ，また学習データ量が限られていたため，過学習を起こしたのだと考えられる．観測対象に対して適したキャパシティを持つようにモデルを使用するべきだと示唆できる．

6. 結論

本研究では，特定の検索行動を示唆する共起語を持つ検索クエリが観測に適していること，また各クエリに対して適したモデルが存在することを示した．これにより，検索クエリを用いた事象観測における素性抽出とモデル選択に関して，共起語に着目した方法に関する知見を得た．

謝辞

本研究は JSPS 科研費 JP25700032, JP15H05327, JP16H06562 の助成を受けたものです.

参考文献

[Sakaki 2012] Sakaki, Takeshi and Matsuo, Yutaka. (2012). Twitter as a Social Sensor : Can Social Sensors Exceed Physical Sensors?(Special Issue: Twitter and Social Media) In Journal of Japanese Society for Artificial Intelligence, 09128085, pages 67–74, The Japanese Society for Artificial Intelligence.

[Choi 2009a] Choi H, V. H. (2009a). Predicting Initial Claims for Unemployment Benefits.

[Choi 2009b] Choi, Hyunyoung and Varian, Hal (2009b). Predicting the present with google trends. In The Economic Record, pages 2-9.

[Goel 2010] Goel S., Hofman J. M., L. S. P. D. M. and J., W. D. (2010). Predicting consumer behavior with Web search. In Proceedings of the Na- tional academy of sciences 107.41 (2010): 17486-17490.

[Goel 2010] Goel S., Hofman J. M., L. S. P. D. M. and J., W. D. (2010). Predicting consumer behavior with Web search. In Proceedings of the Na- tional academy of sciences 107.41 (2010): 17486-17490.

[Polgreen 2008] Polgreen, Philip M. and Chen, Yiling and Pennock, David M. and Nelson, Forrest D. and Weinstein, Robert A. (2008). Using Internet Searches for Influenza Surveillance Clinical Infectious Diseases 47.11 (2008): 1443-1448.