5E-07

検索条件のベクトル表現を用いた検索の興味の可視化

布施拓馬 [1)]
早稲田大学 [a)]

關翼人 [2)]
早稲田大学 [a)]

村田昇 [3)]
早稲田大学 [a)]

杉浦太樹 [4)]
株式会社リクルート住まいカンパニー [b)]

野村眞平 [5)]
株式会社リクルート住まいカンパニー [b)]

1. はじめに

条件設定型の検索，例えば不動産検索では，家賃などの検索条件を複数設定しながら検索を進める．ユーザーの重要視している検索条件を推定できれば，レコメンドへの応用が期待できる．重要な検索条件は過去の検索内で類似した検索条件として現れるという仮定のもと，ユーザーの重要視している条件及び条件の遷移についての推定を行う．本稿では検索条件のモデル化を行い，検索条件の推移を過去の検索と現在の検索の重み付き cross entropy の最小化として定式化する．不動産検索ポータルサイトにおける実際の検索ログデータを用いて検索条件の可視化及びユーザーの検索の興味の可視化を行い，提案方法の検証を行う．

2. 検索条件に対するベクトル表現の獲得

あるユーザーが検索条件を変化させながら検索を進める例を考える．ここでは検索が時系列的に積み重なったものをログデータと呼ぶ．各検索は検索条件の集合と考えることができ，時刻 t の検索における i 番目の検索条件を s_t^i とすると，検索は $\{s_t^i\}_{i=1}^{N_t}$ と表せる．N_t は時刻 t の検索で指定された検索条件の個数を表す．検索条件は間取りなどのカテゴリカルデータや，家賃などの数値表現されていてもその距離関係が人間の感覚と一致しないデータである．そこで単語，文章，文書と検索条件，検索，ログデータのアナロジーを仮定し，word2vec[2] を用いて各検索条件のベクトル表現を獲得する．

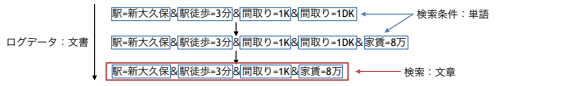

Fig. 1　検索と文書のアナロジー

word2vec では文章内で共起した単語は類似度が高いという仮定を置いているが，検索に関しても同様に，共に設定される検索条件同士は類似度が高いという仮定を置くことができ，ログデータを入力することで検索条件のベクトル表現が獲得できると考えられる．本節以降では検索条件 s_t^i が d 次元ベクトル表現を表すものとする．

Visualization of User Interest based on Vector Representation of Searching Query
1) Takuma FUSE　2) Yokuto SEKI　3) Noboru MURATA
4) Taiki SUGIURA　5) Shimpei NOMURA
a) Waseda University
b) Recruit Sumai Company Ltd.

3. 検索の興味の重み付け

本章では検索を続けるユーザーがどの検索条件を重要視しているかを推定するという問題について論ずる．

3.1 問題設定

検索を検索条件とそれぞれに重要度を表す重みがついたデータ集合として扱い，これを Bag of Data(BoD) と呼ぶ．検索条件の BoD は s_t^i につく重みを $w_t^i \in \mathbb{R}_{\geq 0}$ と置くと，$B_t = \{s_t^i, w_t^i\}_{i=1}^{N_t}$ と表せる．また，どの検索条件を重要視しているか推定するという問題を，BoD 内の検索条件に対する重み最適化として定式化する．現在指定した検索条件の内，重要なものは過去に類似した検索条件として現れるという仮定を置き，現在の検索と過去の検索の重なりを考慮する．本稿では，現在の検索条件の BoD と，過去の検索条件の BoD の多重集合をそれぞれ粒子近似された経験分布とみなし，二つの分布同士の重なりを抽出する．例えば家賃を 3 万円，3.5 万円，4 万円と変えたユーザーを考える．このユーザーは一致した条件を選択した訳ではないものの，安い物件を探している，という興味が伺える．このような例に対し，検索条件の指定された数のカウンティングだけでは推定できない興味が，分布同士の重なりを抽出することで推定できると考えられる．本稿では重み付きデータに対する cross entropy を推定できる Mean Quantile Cross Entropy Estimator(MQCEE)[1] を用い，二つの分布の cross entropy を最小化する重みを推定する．MQCEE は二つの BoD$B_t = \{s_t^i, w_t^i\}_{i=1}^{N_t}$，$B_{t'} = \left\{s_{t'}^j, w_{t'}^j\right\}_{j=1}^{N_{t'}}$ 間に対し，式 (1) のように定義される．

$$H(B_t, B_{t'}) = \log c_d + 1 + d \sum_{i=1}^{N_t} \sum_{j=1}^{N_{t'}} \frac{w_t^i w_{t'}^j \log \|s_t^i - s_{t'}^j\|}{\sum_{i=1}^{N_t} w_t^i \sum_{j=1}^{N_{t'}} w_{t'}^j} \quad (1)$$

式 (1) の c_d はガンマ関数を Γ として $\frac{\pi^{d/2}}{\Gamma(1+d/2)}$ と書ける．

3.2 検索の興味の可視化

過去の検索の BoD を $B^{t-1} = \{B_1, ..., B_{t-1}\}$ と置く．最適化した重みは MQCEE を用いて式 (2) のように得られる．実用上重みの総和が 1 という制約をつける．

$$\arg\min_{\{w_t^i\}_{i=1}^{N_t}} H(B_t, B^{t-1}) \quad \text{subject to} \sum w_i = 1 \quad (2)$$

過去の検索についても重要度を重みとして導入することができる．本稿では検索を続ける上で興味の対象が変遷する場合を想定している．例えばあるユーザーが駅付近を検索していたが，想定より家賃が高く，郊外でも家賃が安い物件を検索した例を考え

る．このユーザーは最初は駅に興味を持っていたものの，途中から賃料に興味が変遷したと言える．そこで，直近のデータほど推定に有用という仮定を置き，B^{t-1} 内で，時系列を考慮した重みを与える．本稿では，経過時間に反比例した重みをつけることを考える．これは直近の時刻の検索条件ほど大きな重みがつくことを意味し，過去の検索の BoD を $B^{t-1} = \{\{s_i^1, \frac{w_i^1}{t-1}\}_{i=1}^{N_1}, \{s_j^2, \frac{w_j^2}{t-2}\}_{j=1}^{N_2}, ..., \{s_k^{t-1}, \frac{w_k^{t-1}}{1}\}_{k=1}^{N_{t-1}}\}$ とすることで時系列を考慮した重みが推定できる．

4. 実験

本章では不動産検索ポータルサイト SUUMO で収集された賃貸物件の検索ログデータの一部を利用し，各検索条件のベクトル表現獲得と検索条件の重みの推定の実験を行う．具体的には約 2000 個の検索条件から構成される 540 万件の検索を用いて実験を行った．

4.1 検索条件のベクトル表現獲得について

ログデータから，最寄り駅や間取りなど 14 種の条件カテゴリを選び，それらの具体的な検索条件約 2000 個に対して 100 次元のベクトル表現を獲得した．そのうち最大賃料と間取りのベクトル表現の可視化を Fig.2, Fig.3 に示す．可視化には tSNE[3] を用いた．

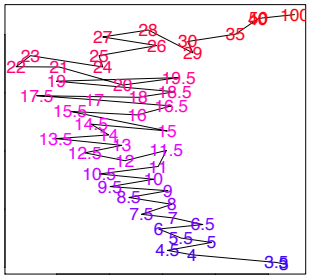

Fig. 2 最大賃料の可視化

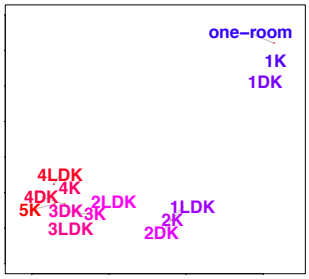

Fig. 3 間取りの可視化

Fig.2 からは共起性からベクトル表現を獲得したにも関わらず，数値的な関係性が確認できる．また，Fig.3 からは三つのクラスタが確認できる．これは賃貸を借りる人は，一人で住む人，二人で住む人，三人以上で住む人に分けられやすい，という不動産的な知見と一致する．

このように Fig.2，Fig.3 からは，数値的順序関係を維持し，かつ不動産を利用するユーザーの特性を内包した検索条件のベクトル表現を獲得できたことが分かる．

4.2 検索条件の重みの推定

検索条件のベクトル表現を用いて，あるユーザーの興味の推定を行った．本稿では視認性を上げるため 14 種の条件カテゴリの重みを推定し，可視化した．また，最寄り駅等，複数の条件を設定できる条件カテゴリに対しては複数の重みの内，最大値を代表値としてその条件カテゴリに対する重みとした．54 回検索を行ったユーザーの重みの推定結果を Fig.4 に示す．横軸が検索回数，縦軸が今回ユーザーに用いられた 6 種の条件カテゴリを表しており，各時刻における円の大きさが重みの大きさと対応しているため，大きな円はその条件カテゴリを重要視していることを表している．なおその条件を指定せずに検索を行った場合に円は書かれない．また，ログデータの変化のタイミング 5 つに関してその内容を記載した．

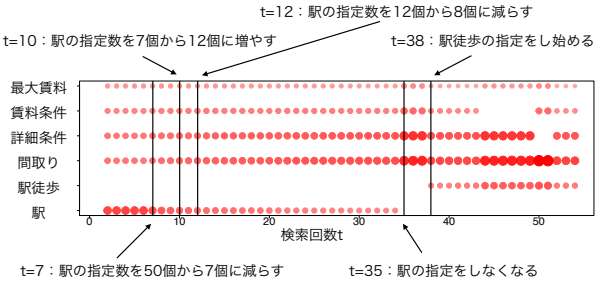

Fig. 4 検索条件の重みの推定

このユーザーは最初は駅に着目し検索を進めていたが，次第に間取りなど駅以外の検索条件に興味が遷移していったことが読み取れる．t=35 において駅に関する検索条件を外しているが，t=35 以前に駅に対する興味が失われていく様子が確認できる．また，t=10 で指定している駅の数を増やしているにも関わらず，駅に対する興味は増加していない．これは新しく追加された駅が今まで興味を抱いていた地域や路線と違うものであり，今まで検索していた駅に対する興味の向上の根拠にならないとみなされたと考えられる．実際に直後の t=12 では駅の指定数が 8 個に減少している．

5. まとめと今後の展望

本稿では検索条件に対するベクトル表現を獲得し，重み付き cross entropy の最小化を用いて検索の興味の可視化を行った．結果としてある検索条件に対する興味が失われていき，興味が変遷する様子が確認できた．今後の展望としては，獲得した検索条件や検索の数値表現を元にした各ユーザーに対する最適な検索や検索条件のレコメンデーションの実現が想定される．

参考文献

[1] H. Hino and N. Murata “Information estimators for weighted observations.” Neural Networks, vol.46, pp.260 - 275, 2013.

[2] T. Mikolov, K. Chen, G. Corrado and J. Dean “Efficient estimation of word representations in vector space.” CoRR, abs/1301.3781, 2013.

[3] L.J.P. van der Maaten and G.E. Hinton. “Visualizing High-Dimensional Data Using t-SNE.” Journal of Machine Learning Research 9(Nov):2579-2605, 2008.