C-15

滞在時間と遷移確率による類似ユーザの判別と，対象ユーザの移動先予測

Distinction of Similar Users and Intended User's Moving Projection Using Stay Time and Transition Probability Density

中村　亮太†　　THAWONMAS Ruck‡
Nakamura Ryota　　Ruck Thawonmas

1．はじめに

近年ＧＰＳなどの位置情報取得技術が発展しており，ユーザの現在位置を用いたサービスを提供するシステムが提案されてきている．このようなシステムは location-aware システムと呼ばれており，現在位置から近い位置にある店や観光地などを表示する街案内システムも，情報取得機能を持った PDA(携帯情報端末)や携帯電話の普及により,様々な用途で利用されるようになってきた．

しかし，殆どの既存のシステムでは予測に学習用のユーザ全体を用いるものが多い．学習用のユーザから類似ユーザを抽出し予測に適用すれば精度が良くなると考えられる．また，サービスの対象とするユーザの滞在した時間や位置情報の履歴を活用せずに移動先の予測や推薦システムの実装を行っている．対象ユーザが現時点までに辿った位置情報と各々に滞在した時間を利用することで，予測精度が良くなり，遷移確率を用いた推薦システムでも性能の改善が見込まれるはずである．そこで，本論文ではコンテンツの滞在時間とコンテンツ間の遷移確率の 2 つの観点から，それぞれ対象ユーザと類似するユーザを判別する．そして，類似ユーザの情報を用いて行列分解法とバックオフスムージングにより対象ユーザの移動先を予測する手法を提案する．

2．関連研究

過去のlocation-aware システムに竹内らの位置情報履歴を利用したユーザアダプティブな街案内システム[1]がある．このシステムでは協調フィルタリングの技術とエリア分割したマップにおける遷移確率を利用して，ユーザの好みに合う店を推薦する．

博物館における移動先予測では，Bohnert らの展示物閲覧時間と遷移確率を組み合わせる手法[2]と類似ユーザの情報と遷移確率を組み合わせる手法[3]がある．前者の手法では類似ユーザの展示物閲覧時間を確率化したものと，全ユーザの遷移確率を組み合わせ，最終的な確率として移動先を予測している．後者の手法では類似ユーザの数に類似度による重みをかけたものを確率化したものと全ユーザの遷移確率を組み合わせ，最終的な確率として移動先を予測している．

どの手法においても，確率の計算において類似ユーザのみのデータを用いていない．また，遷移確率を求める際に対象ユーザの現地点での位置情報しか用いておらず，対象ユーザの履歴情報を用いていない．

3．提案手法

類似ユーザの判別と行列分解法，バックオフスムージングを用いた予測手法について提案する．本手法では精度を向上させるために一度滞在したコンテンツにもう一度滞在することはないとして扱う．また，類似度計算や予測のために学習用ユーザは 3 つ以上のコンテンツに滞在しているデータを用いる．対象ユーザも 3 つ以上滞在した時点から計算を行うものとする．

3.1　類似ユーザの判別

滞在時間と遷移確率の 2 つの観点から類似ユーザを判別する．滞在時間からは対象ユーザとコンテンツに滞在する時間が類似するユーザを抽出する．遷移確率からは対象ユーザと遷移の履歴が類似するユーザを抽出する．

3.1.1　滞在時間による類似ユーザの判別

滞在したコンテンツの評価値を以下式により計算する．

$$R_{ui} = \frac{t_{ui}}{\bar{t}_{u\bullet}} - \frac{1}{x_{\bullet i}} \sum_{v \in U} x_{vi} \frac{t_{vi}}{\bar{t}_{v\bullet}} \cdots (1)$$

u は対象とするユーザ，U は学習ユーザ群，v は学習ユーザ，i は対象のコンテンツ，●は全てのコンテンツ又は全てのユーザである．また，x は i に滞在した場合 1，していない場合は 0 となる．

（1）式の右辺左側の項は対象とするユーザの着目しているコンテンツを，今までに滞在したコンテンツの平均と比較している．右側の項は着目しているコンテンツに滞在した他のユーザに同様の計算を行い，平均をとっている．この式により対象ユーザ自身と他のユーザの両方を比較した評価値を得ることが出来る[2]．

得られた評価値から以下式により，対象ユーザとの類似度を計算する．

$$Sim\_time(u,v) = \frac{\sum_i R_{ui} R_{vi}}{\sqrt{\sum_i {R_{ui}}^2} \sqrt{\sum_i {R_{vi}}^2}} \cdots (2)$$

類似度が閾値を越えるユーザ郡を滞在時間による類似ユーザとする．

†立命館大学大学院．修士生．Ritsumeikan University graduate school master student.

‡ 立命館大学知能情報学科．教授．Professor Ritsumeikan University intelligence information subject.

3.1.2 遷移確率による類似ユーザの判別

対象ユーザが滞在したコンテンツの総数を N_u とする. また, 対象とするユーザが最後に滞在したコンテンツに着目し, 学習用のユーザの遷移履歴においてそのコンテンツまでの履歴を抽出し, そのコンテンツの総数を N_v とする. 対象ユーザの履歴と抽出した学習ユーザの履歴を比較し, コンテンツ間の同一遷移数を数え, N_{trans} とする.対象ユーザが最後に滞在したコンテンツが含まれていない学習ユーザとの類似度は0として扱う. また,

得られた変数より, 以下式で類似度を計算する.

$$Sim\_trans(u,v) = 1/(N_u - N_{trans})(N_v - N_{trans}) \ldots (3)$$

類似度が閾値を越えるユーザ群を遷移確率による類似ユーザとする.

3.2 行列分解法

行列分解法は n 人のユーザによる m 個のコンテンツの評価値行列 $R(n \times m)$ を, よりランクの低い 2 つの行列 $P(n \times k)$ と $Q(k \times m)$ との積で近似する手法である.

本論文では P の各行ベクトルを p_u , Q の各列ベクトルを q_i とし, 以下式を満たすような P, Q を確率勾配法（最急降下法）によって求める.

$$\min_{q_\bullet, p_\bullet} \sum_{(u,i)\in K} (R_{ui} - q_i^T p_u)^2 + \lambda(\| q_i \|^2 + \| p_u \|^2) \ldots (4)$$

λ は正規化の重みである. $R(n \times m)$ に欠損値が多いため正規化項を付けて最適化する必要がある. 予測誤差を

$$e_{ui} \overset{def}{=} R_{ui} - q_i^T p_u$$

とし, 以下式で p と q を更新していき最適解を求める.

$$q_i \leftarrow q_i + \gamma(e_{ui} p_u - q_i)$$

$$p_u \leftarrow p_u + \gamma(e_{ui} q_i - p_u)$$

γ は学習率である.

3.1.1 で求めた類似ユーザを用いて, 行列分解法によって対象ユーザの評価値行列 $R_{u\bullet}$ の欠損値を補完する. 補完した 評価値を[0,1]の範囲に正規化し, 確率 p_{time} とする.

3.3 バックオフスムージング

N-gram モデルは N 単語連鎖の統計に基づいて, （N-1）単語の履歴から次の単語の生起確率 p を与えるものである.

$$p(w_k \mid w_{k-N+1}, \ldots, w_{k-1}) = \frac{C(w_{k-N+1}, \ldots, w_{k-1})}{C(w_{k-N+1}, \ldots, w_k)} \ldots (5)$$

w は単語の履歴, C は頻度の計数である.

N=1 の場合を unigram, N=2 の場合を bigram, N=3 の場合を trigram と呼び, 多くの場合では trigram が効果的といわれている. N-gram モデルの信頼性を向上させるために, 出現数の少ない単語列を学習から削除する手法や, 確率が 0 となるのを防ぐためのスムージング手法が提案されている. 精密なスムージングとして, 単語履歴に応じて, trigram があればそれを補正して用い, なければ bigram に係数を乗じて利用する（それもなければ unigram を用いる）方法があり, バックオフスムージングと呼ばれ, 以下のように確率を推定する.

$$\begin{cases} \alpha * p(w_k \mid w_{k-2}, w_{k-1}) \ldots \text{trigramが存在} \\ \beta * p(w_k \mid w_{k-1}) \ldots \text{履歴のみ存在} \\ p(w_k \mid w_{k-1}) \ldots \text{履歴が存在しない} \end{cases}$$

w は単語の履歴である. また, α をディスカウント係数, β をバックオフ係数と呼ぶ. α の推定法にはいくつかの方法が提案されている. Witten-Bell 法では,

$$\alpha = \frac{\text{trigramの総数}}{\text{trigramの総数} + \text{trigramの種類数}}$$

としている. β は履歴毎に確率の挿話を 1 にするような正規化係数である[5].

本手法では単語をコンテンツに置き換え, 遷移の履歴を活用する. 3.1.2 で求めた類似ユーザを用いて, バックオフスムージングにより遷移確率 p_{trans} を求める.

3.2 で求めた p_{time} と 3.3 で求めた p_{trans} を以下式で組み合わせる.

$$p(i) = \omega p_{time}(i) + (1-\omega) p_{trans}(i) \ldots (6)$$

ω は p_{time} と p_{trans} のどちらの比率を大きくするかの重みである.

4. おわりに

本論文では滞在時間と遷移確率の 2 つの観点から, それぞれ類似ユーザを判別した. また, 類似ユーザを用いた行列分解法とバックオフスムージングにより, 対象ユーザが次に移動するコンテンツの遷移確率を求める手法を提案した. 今後の展望としては提案手法の性能を確かめるために, 既存手法と比較実験を行い, 有用性を検証する.

参考文献

[1] 竹内雄一郎, 杉本雅則. 位置情報履歴を利用したユーザアダプティブな街案内システム. 電子情報通信学会論文誌 D, Vol.J90-D, No.11, pp.2981-2988, 2007

[2]Fabian Bohnert, Ingrid Zukerman, Shlomo Berkovsky, Timothy Baldwin, and Liz Sonenberg: "Using Interest and Transition Models to Predict Visitor Locations in Museums", Technical report 2008/219, Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia, 2008.

[3]Fabian Bohnert and Ingrid Zukerman: "Personalised Pathway Prediction". In Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization (UMAP-10), pages 363–368, Waikoloa, HI, USA, 2010.

[4] Yehuda Koren; Robert Bell; Chris Volinsky,"Matrix Factorization Techniques for Recommender Systems". IEEE Computer, 2009, 8

[5]河原達也. 音声言語モデル http://www.ar.media.kyoto-u.ac.jp/~kawahara/paper/lm.pdf