H-048

SimRankによる類似POI検索に関する検討

A Study on POI Similarity Search via SimRank

黒川 茂莉† 横山 浩之†
Mori Kurokawa Hiroyuki Yokoyama

1. まえがき

GPS などによる測位環境が整い，携帯電話端末を使った位置情報の履歴の収集が容易になった．この位置情報履歴を分析することにより，自宅，会社，デパート，レストランなど，さまざまな意味や役割をもつ場所へ滞在した事象が抽出できると考えられる．ここで意味や役割をもつ場所を Points of Interest (POI) と呼び，ある位置情報の範囲を代表する名前として表現されるものとする．本論文では主として店舗名を POI とする．

我々は，位置情報履歴の分析の結果得られるであろう多数のユーザの POI への滞在履歴を使って，個々のユーザの行動傾向に類似した POI の検索を行う方式を検討している．本論文の類似 POI 検索は，POI をアイテムとし，ユーザの POI への滞在をユーザのアイテムに対する評価とみなすことで，ユーザのアイテムに対する評価履歴を基にしたユーザへのアイテムの推薦という枠組でとらえることができる．

評価履歴を基に対象ユーザに対する推薦を行う方式として，協調フィルタリング (Collaborative Filtering: CF) [2] があり，ユーザベースの方式とアイテムベースの方式がある．ユーザベースの CF は，履歴が類似したユーザは同じアイテムを評価する傾向が強いと仮定し，推薦対象ユーザの推薦候補アイテムに対する予測評価を，推薦対象ユーザと類似度が高いユーザの候補アイテムに対する評価がより強く反映するように計算する．一方，アイテムベースの CF は，履歴が類似したアイテムは同じユーザに評価される傾向が強いと仮定し，推薦対象ユーザの推薦候補アイテムへの評価を，推薦候補アイテムと類似度が高い既評価アイテムへの評価がより強く反映するように計算する．

CF を POI の推薦に適用した場合，以下の課題への対応が必要であると考えられる．

1. 暗黙的評価：POI に滞在した場合のみ POI への評価として収集するので，unary (1 値) の評価しか得られない．
2. リアルタイムな推薦要求：検索と同時に推薦要求が行われることが想定されるため，応答速度が求められる．
3. スパースネス：ユーザ・POI の組み合わせに対して滞在数が少なく，適切な推薦が難しい．このように評価行列がスパースな場合には，Top-N 推薦の適合率が低下することが指摘されている[2].

とくに，アイテムを POI ととらえた場合に 3 番目のスパースネスの問題の影響は大きく，改善の効果は大きいと考えられる．

以降，2 章では提案手法，3 章では評価実験およびとその結果，4 章ではまとめを示す．

2. 提案手法

我々の提案手法は CF を基礎とし，1 章で述べた 3 つの課題それぞれに対する対策を行った．1 番目の暗黙的評価への対応として，unary 評価ありの場合を 1，unary 評価なしの場合を 0 とし，2 値評価として扱う．2 番目のリアルタイムな推薦要求への対応として，アイテムベースの CF とする．アイテムベースとすることによりアイテム類似度の事前計算が可能となり，リアルタイムな推薦要求への応答速度が向上する．最後に，3 番目のスパースネスへの対応として，SimRank[1] と呼ばれるリンク解析により類似度計算を行う．SimRank は，文献などのノード間の参照 (引用) 関係から文献間の類似度を計算する目的で用いられるリンク解析手法で，大量のノードに対して少数のリンクしかないスパースな参照行列に対して有効である．さらに，我々は POI 間の移動が多いほど POI 間の関係は強いと仮定し，SimRank に対して POI 間の移動回数をより強く反映する拡張 (以降，「リンク拡張」と呼ぶ) を行った．

以下，提案手法における具体的な予測評価の計算方法および類似度の計算方法を説明する．

アイテムベースの推薦とするため，対象ユーザの候補 POI に対する予測評価は，既評価 (滞在) POI に対する評価を候補 POI と既評価 (滞在) POI の類似度で重み付けした加重和により計算される．具体的には，対象ユーザを U，候補 POI を A，対象ユーザ以外の任意の既評価 POI を B_i とすると，対象ユーザの候補 POI に対する予測評価 $p(U,A)$ は次式で計算される．

$$p(U,A)=\sum_i sim(A,B_i)v(U,B_i)$$

ここで，$v(U,B_i)$ は，対象ユーザ U の既評価 POI B_i に対する評価値である．また，$sim(A,B_i)$ は POI A,B_i 間の類似度であり，これを SimRank により予め計算する．

SimRank をユーザの POI への滞在履歴のような二部グラフに適用した場合，ユーザ間の類似度と POI 間の類似度が相互に伝搬するように繰り返し計算を行う．具体的には，U,V をユーザ，A,B を POI とし，$I(X)=\{I_i(X);i=1,\cdots|I(X)|\}$ をノード X へのインリンク (入ってくるリンク)，$O(Y)=\{O_i(Y);i=1,\cdots|O(Y)|\}$ をノード Y からのアウトリンク (出ていくリンク) とし，$L(A,B),L(B,A)$ をそれぞれ A から B，B から A へのリンク数とし，C を減衰係数 (decay factor) とすると，ユーザ間の類似度 $sim(U,V)$，POI 間の類似度 $sim(A,B)$ は次式で計算される．$sim(A,B)$ の分母，分子の項 $(L(A,B)+L(B,A))^2$ がリンク拡張部分である．

$$sim(U,V)=\frac{C}{|O(U)||O(V)|}\sum_{i=1}^{|O(U)|}\sum_{j=1}^{|O(V)|}sim(O_i(U),O_j(V))$$

$$sim(A,B)=C\frac{\sum_{i=1}^{|I(A)|}\sum_{j=1}^{|I(B)|}sim(I_i(A),I_j(B))+(L(A,B)+L(B,A))^2}{|I(A)||I(B)|+(L(A,B)+L(B,A))^2}$$

上式の計算により，POI A,B に対するインリンクの始点となるユーザ間の類似度が POI A,B 間の類似度に反映される．

†（株）KDDI 研究所 KDDI R&D Laboratories, inc.

ここで，類似度の初期値は，同一ユーザ，同一 POI の類似度を 1 とし，その他を 0 とする．また，減衰係数 C は 0.8 とする．

3．評価実験およびその結果

提案手法の評価のため，2009/1/17～2009/2/8 に，週末におけるユーザの POI 滞在履歴をアンケートにより取得した．ユーザには期間中に，新宿エリア，渋谷エリア，銀座・有楽町エリアのうち 2 エリア，各 2 回の行動を必須とし，1 回の行動について 3 スポット以上を回ることを必須とした．データの内容は以下の通りである．

- ユーザ：212 人
  - 20-59 歳の男女 (男性　106 人　女性　106 人)
  - 2 つ以上のエリアに 2 週間に 1 回以上滞在すると回答したユーザからランダムに選定した．
- POI：3,574 店舗
  - 新宿エリア　1,613 店舗　渋谷エリア　1,023 店舗　有楽町・銀座エリア　938 店舗
  - 実店舗との対応づけがとれなかったものは，個別の店舗とした．
- 延べ滞在 POI 数：4,851
  - 1 週目　1,264　2 週目　1,268　3 週目　1,258　4 週目　1,061

下図は，ユーザの POI への滞在頻度を図示したものであり，縦軸がユーザ，横軸が POI，黒い点がユーザの POI への 1 回以上の滞在を表す．

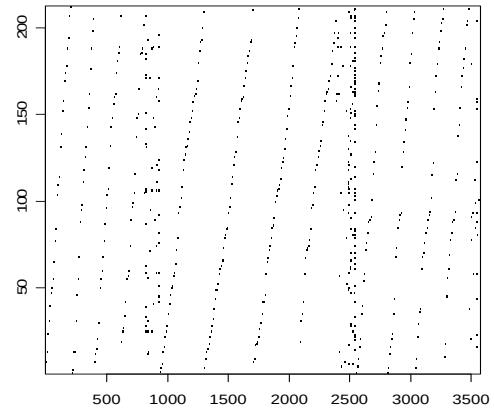

図1　ユーザの POI への滞在頻度

評価実験における推薦シナリオは，その日にユーザが滞在しているエリアとその日のそれまでの行動経路上の POI が分かっている場合に，次に滞在する可能性が高い店舗 POI を推薦するものとする．3 週目までの滞在履歴を学習データとし，4 週目の滞在履歴を検証用データとする．学習データに含まれないユーザおよび実店舗との対応づけがとれなかった POI は検証用データから除外した．

リンク拡張なし，リンク拡張ありのそれぞれの場合について，反復回数を 1 回～5 回と変え，提案手法による Top-N 推薦を行った．[2] の実験と同様に，各条件での Top-N 適合率 (%) を計算した結果は，表 1，2 の通りである．Top-N 適合率とは，各検証用データについて正解 POI を見ずに Top-N 推薦を行った場合に，推薦結果に正解 POI が含まれる割合である．

表1　Top-N 適合率 (リンク拡張なし)

	1回	2回	3回	4回	5回
Top-10	0.00	5.08	10.66	10.66	10.66
Top-20	0.00	9.64	32.99	34.52	34.01
Top-50	0.00	10.66	55.84	56.35	56.35
推薦不能	100	89.34	34.52	34.01	34.01

表2　Top-N 適合率 (リンク拡張あり)

	1回	2回	3回	4回	5回
Top-10	0.00	12.18	15.23	15.74	15.23
Top-20	0.00	20.81	38.07	38.58	38.07
Top-50	0.00	23.35	60.41	60.41	60.41
推薦不能	100	76.14	33.50	32.99	32.99

反復回数=1 回の場合は，単純なベクトル類似度に基づくアイテムベース CF の場合に相当するが，表 1 によれば，この場合には Top-100 推薦 POI の中に正解 POI が含まれることはなかった．また，反復回数を増加させると，適合率が上昇し，反復回数=4 回以上の場合の適合率はほぼ一定となった．これにより，反復回数=4 回で，類似度の値が収束していることが分かる．

リンク拡張なし (表 1) とリンク拡張あり (表 2) との場合を比べると，リンク拡張ありの場合は反復回数が少ない場合や Top-10 推薦，Top-20 推薦の場合に適合率が高いことが分かった．

なお，推薦不能な POI の割合が高い理由としては，少数のユーザしか滞在していない POI が存在するため，4 週目の行動経路上の POI と推薦候補 POI の類似度が計算できない場合があることが考えられる．

4．まとめと今後の課題

ユーザの POI への滞在履歴を用い，SimRank を使った POI 間の類似度を算出し，類似検索を行う手法を提案し，推薦精度の観点で有効性を評価した．評価の結果，SimRank を用いる方法がスパースネスの問題に有効であり，3 章で述べた SimRank に対するリンク拡張は少ない反復回数の場合や Top-10 推薦，Top-20 推薦の場合に有効であることを確認した．

粒度の異なる POI の推薦に対応すること，推薦精度を実用的な精度に高めることが今後の課題である．また，本研究と同じ枠組みでとらえたナビゲーションに関する研究として[3], [4] があり，これらと精度評価やシステム評価の観点で比較することも今後の課題である．

参考文献

[1]Glen Jeh, Jennifer Widom, “SimRank: a measure of structural-context similarity”, Proc. KDD 2002, pp.538-543, (2002).
[2]Matthew R. McLaughlin, Jonathan L. Herlocker, “A collaborative filtering algorithm and evaluation metric that accurately model the user experience”, Proc. SIGIR 2004, pp.329-336, (2004).
[3]篠田 裕之, 竹内 亨, 寺西 裕一, 春本 要, 下條 真司, “行動履歴に基づく協調フィルタリングによる行動ナビゲーション手法”, 情報処理学会研究報告, GN, 2007(91), pp.87-92, (2007).
[4]篠田 裕之, 竹内 亨, 寺西 裕一, 春本 要, 下條 真司, “ユビキタス環境における協調フィルタリングを用いた行動ナビゲーション手法の考察”, 情報処理学会研究報告. CSEC, 2007(16), pp.77-82, (2007).