別紙 2

論文審査の結果の要旨

論文提出者氏名　西岡賢一郎

本論文は位置情報 SNS (Social Networking Servicce) のチエックインデータ中の位置 (座標) の情報から、ユーザの時刻ごとの空間分布を推定し、そこから疎密の周期性や突発的集中などのパターンを抽出し、それらの手法と知見を将来の位置の予測に活かす方法を研究したものである。

第 1 章は導入である。位置情報 SNS のデータでは、位置データが連続値でかつサンプリングの時刻がユーザごとに任意のタイミングで起こり、しかも 1 ユーザあたりの平均チェックイン回数が 1 日 1 回程度しかないというデータの特性と、そのような疎なデータを使用して 1 日単位よりも細かい分析をする際の難しさを説明している。第 2 章で先行研究を説明している。

第 3 章では、2 次元の拡散方程式に基づき、チェックインの間のユーザの空間分布 $P(\boldsymbol{x}|t)$ を与えている。次に、$P(\boldsymbol{x}|t)$ を空間成分 $U_m(\boldsymbol{x})$ と時間成分 $V_m(t)$ の重み ν_m の積和の $P(\boldsymbol{x}|t) = \sum_{m=1}^{M} \nu_m U_m(\boldsymbol{x}) V_m(t)$ に分解する。この分解を時刻 t と時刻 t' の空間分布の共分散行列 $\Sigma(t, t')$ の閉じた式を与えてから時間に関して離散化し、PCA (Principal Component Analysis; 主成分分析) と ICA (Independent Component Analysis; 独立成分分析) を用いて求める方法を示している。次に得られた空間分布から空間・時間成分を抽出し、時間成分からは FFT により周波数成分を取り出している。全ユーザの空間分布の平均を使って分析した結果からは、定常的に都心部の存在確率が高い、1 日や半日周期で都心部の存在確率が高くなり、1 週間周期で都心部の人の存在確率が低下する、突発的な存在確率の低下 (確認できる事象としてはデータの欠損) があるなど、現象と対応すると思われるものが確認されている。

この結果は、チェックインデータのように疎なデータでも、多数のユーザに対するものを使用することで意味のある空間・時間成分が抽出できることを示しており、提案されたモデルは意味のあるものであると本審査委員会は判断した。

第 4 章では、指定された時刻の空間分布 $P(\boldsymbol{x}|t)$ を推定する時に使用しているスケールパラメータ λ を求める手法を提案し、第 3 章で提案した手法と比較している。第 3 章で提案した λ の最尤推定は 3 つの連続するチェックインに対して、前後で決まる空間分布のもとで、途中のチェックインの位置の確率が平均して最大となるように最尤推定を行う。しかしこの推定手法では同じ時刻に 2 箇所にチェックインしてしまった場合などチェックインデータに異常があるとそれが推定に大きな影響を与えてしまい、異常に小さな λ の値を与えてしまうことがある。これを防ぐため、第 4 章では、事前分布を与えることで異常なデータの影響を抑える推定手法を提案している。再サンプリングしたチェックインデータに適用することで、最尤推定で推定した λ はサンプルにより値が大きく変動するが、事前分布を用いて推定した λ は安定であることが示されている。また、交差検定を用いてチェックインデータの対数尤度の平均値を様々な λ に対して求めたものが最大値を取る時の λ が事前分布を用いて推定した λ に近く、最尤推定を用いて求めた λ とはかなり外れていることも示されている。

λ の正確な値を求める必要があるわけではないが、交差検定よりも簡単に適切な範囲の値を求める手法を提案しており、事前分布を用いた λ の推定手法には一定の意味があると審査委員会は判断した。

第5章では、ユーザのチェックインデータから空間分布を推定する時に、類似したユーザのチェックインデータだけを平均することで、ユーザの空間分布の推定の精度をあげ、より詳細な分析ができる手法を提案している。類似したユーザを決めるためのユーザ間の距離としてはヘリンジャー距離を用いている。確率分布の間の距離の定義にはいくつかあるが、ヘリンジャー距離はユークリッド距離の条件を満たしているため、凝集法のクラスタリングのアルゴリズムの中で性能が高いと言われている Ward 法を適用できる利点がある。指定した時刻の空間分布間のヘリンジャー距離は閉じた式で与えられる。これをユーザ間の距離として使用するためには時間について平均する必要があるが、これは時間について離散化して平均化することで求めている。実験ではユーザを 10 個のクラスタに分類し、第3章の方法を用いてそれぞれのクラスタに属すユーザのチェックインデータから空間・時間成分を抽出した。その結果から、お台場のコミックマーケットに年末に集まる突発的なパターンや東京・横浜に周期的に行くパターンなどを発見している。

ヘリンジャー距離自体はすでに提案されたものであるが、その距離を用いてクラスタリングすることでより精細な分析が可能な手法を提案したことは評価できると審査委員会は判断した。

第6章では、最も新しいチェックインデータ以後の位置の予測のモデルを提案している。第3章で導入した位置の推定を使って最も新しいチェックインデータ以後の位置を推定すると、位置の確率分布は時間とともに広がり続けてしまう。第3章と第5章の実験結果からは1日周期が強く観測されたので、それに基づいて予測する DPM (Diffusion-type Periodic Model)、第5章の方法と組み合わせて類似したユーザのチェックインデータを利用する DPMU (Diffusion-type Periodic Model with similar Users)、更に第3章の方法と組み合わせて寄与度の低い空間・時間成分を省いた空間分布を利用する RDPMU (Reduced Diffusion-type Periodic Model with similar Users) を提案している。ガウス分布と先行研究の PMM と提案手法を、対数尤度と平均順位と平均逆順位で評価したところ、類似したユーザとして適切な数のユーザのチェックインデータを利用した場合、DPMU は先行研究よりも精度が良く、RDPMU よりも計算量が小さく実用的であると結論づけられている。

過去のデータだけでは将来の位置の予測は難しく、実用的にはツイートの内容など他の情報を総合する必要があるが、空間分布だけからでも将来の位置の予測が少しはできることを示した点は評価できると審査委員会は判断した。

第7章は結論と将来への展望である。

以上のように本論文は、位置情報 SNS の位置推定を研究の主な対象とし、空間分布の推定、空間・時間成分への分解によるパターンの抽出、周期性を利用した将来の位置の予測の手法を提案したものであり、位置情報 SNS の分析手法に関連する研究成果として高く評価することができる。したがって、本審査委員会は博士 (学術) の学位を授与するにふさわしいものと認定する。