様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１３　日現在

機関番号：　３２６０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０２６１

研究課題名（和文）確率過程に対する精度保証付き予測シミュレーション法の構築とその知識発見への応用

研究課題名（英文）Developing Predictive Simulation Framework with Confidence Level for Stochastic Process and Its Application to Knowledge Discovery

研究代表者

大原　剛三（OHARA, Kouzou）

青山学院大学・理工学部・准教授

研究者番号：３０２９４１２７

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究では，大規模データのダイナミクスをモデル化する確率過程として，社会ネットワーク上の情報拡散モデルを主な対象とし，限定された試行回数の下でもその確率過程に対するシミュレーション結果の精度を保証する新たな予測シミュレーションパラダイムを統計的機械学習におけるリサンプリング法に基づいて提案し，その有用性を実験的に示した．また，その枠組みをノード中心性指標の計算に応用し，サンプリングした一定数のノードのみから高い中心性指標値をもつノード群を精度よく同定することに成功した．さらに，情報拡散におけるノード影響度を計算する予測シミュレーションを効率よく実行する並列分散アルゴリズムを提案した．

研究成果の概要（英文）：In this work, we addressed a problem of efficiently estimating the influence of a node in information diffusion over a social network. Since the information diffusion is a stochastic process, the influence degree of a node is quantified by the expectation, which is usually obtained by very time consuming many runs of simulation. We proposed a framework for predictive simulation based on the leave-N-out cross validation technique that estimates the approximation error of the influence degree of each node without knowing the true influence degree. We experimentally showed that it can serve as a good measure to solve the problem with far fewer runs of simulation ensuring the accuracy. Besides, we applied that framework to computation of node centrality in order to show the broad utility of the proposed resampling-based framework. In addition, we also devised an efficient algorithm that runs an individual information diffusion simulation in parallel on a distributed computing environment.

研究分野：　データマイニング、社会ネットワーク分析

キーワード：　予測シミュレーション　確率モデル　機械学習　統計数学　知識発見

１．研究開始当初の背景

今日，Facebook や Twitter などのソーシャルメディアの爆発的な普及により，インターネット上に巨大な社会ネットワークが形成されつつある．一旦，ソーシャルメディアに投稿された情報は，ときには複数のメディアを跨ぎつつ，それらの上に展開されている社会ネットワークを通して急速，かつ広範囲に拡散され，非常に多くの人々に共有され得る．そのよう情報，およびその拡散現象は，我々の日常における意思決定にも多大な影響を与えることから，近年，社会学のみならず計算機科学も含めた多様な分野において盛んに研究されており，巨大ネットワーク上を伝播する膨大な情報とその影響を効率的に分析することが急務になりつつある．

そのような社会ネットワーク上の情報拡散に関する研究の中で，特に重要と考えられるのは，情報の拡散を最大化するノード集合を同定する影響最大化問題に関するものである．この問題は，いわゆる口コミマーケティングなどに応用され得るものであり，情報拡散を最小化するために情報の伝播を阻止すべきリンク集合を求める問題やネットワークの外部から新たな情報を投入する際に，その拡散を最大化するような目標ノード集合を同定する問題などは，いずれもこの問題の亜種とみなせる．

一方，これらの問題を解くためには，ノードの影響度を推定し，それに基づいてノードを順位づけることが必要となる．ここで，情報拡散プロセスは確率過程であるため，ノード v の影響度は，v を情報源とした情報拡散プロセス終了時においてその情報を受け取るノード数の期待値として定義される．この考えの下，Kempe らは多数の情報拡散シミュレーション結果の平均を用いてノードの影響度を近似し，貪欲探索に基づき影響最大化問題の解となる最適なノード集合を選択する手法を提案している．しかし，影響度を精度よく近似するためには膨大な回数のシミュレーション試行が必要となり，多大な計算時間を要する．そのため，効率を改善するための様々な手法が提案されているが，シミュレーション試行回数を本質的に削減するには至っていない．

２．研究の目的

以上のような背景の下，本研究では，大規模データのダイナミクスをモデル化する確率過程を主な対象に，限定された試行回数の下でもその確率過程に対するシミュレーション結果の精度を保証する新たな予測シミュレーションパラダイムを機械学習の枠組みで確立することを目的とする．具体的には，社会ネットワーク上の情報拡散モデルを主な対象とし，統計的機械学習におけるリサンプリング法などの理論を基礎として，予測シミュレーションの精度を保証する理論的枠組の構築，およびそれに基づくシミュレーション法の実現を試み，その有用性を検証する．

３．研究の方法

（1）実データの収集と確率モデルの妥当性の検証

本研究では，申請者らのこれまでの研究資産を最大限活用して効率的に研究を進めるため，主な対象データとしてソーシャルデータ（社会ネットワーク）を用いた．具体的には，提案法の評価データとして用いるため，予備的研究で用いていた社会ネットワークに加え，Twitter からより大規模な相互参照ネットワーク，およびそのネットワーク上の情報拡散データを広範囲に収集した．新たに収集した実データに関しては，その統計的性質を分析するとともに，用いる確率モデルが対象データに対して妥当かどうかを確認した．

（2） シミュレーションの予測誤差を精度保証する理論的枠組の検討

本研究では，ある確率モデルに基づく $|S|$ 回のシミュレーションから得られる数値集合 S から，確率変数となるそのサンプル平均を求め，その真の確率分布 $p(x)$ の平均値 μ に対する予測誤差を精度よく推定する問題を考える．このとき，理論的基礎として中心極限定理を利用する．このような前提の下，本研究では，予測誤差を精度保証する学習理論の枠組として，$|S|$ 回の試行サンプルから一定数のサンプルを抽出して統計的な評価をするリサンプリング法に基づいた精度予測を主に検討した．

（3）リサンプリング法に基づく予測シミュレーション法の汎用性の検討

上記（2）で検討する手法は，基本計算を膨大な回数反復実行して得られる平均値をより少ない試行回数で一定精度の下で予測するものであり，この計算原理は本研究で対象とする大規模社会ネットワークにおけるノードの重要性を定量的に表す中心性指標の計算にも共通する．本研究では，その点に着目し，提案するリサンプリング法に基づく予測計算の枠組みをノード中心性の計算にも応用し，その枠組みの汎用性を示すことを試みた．

（4）予測シミュレーション法の並列分散実行環境の構築

社会ネットワークにおけるノード影響度の計算には，これまで述べてきたように膨大な回数のシミュレーションの実行が必要である．本研究で提案する予測シミュレーション法の評価においても，比較すべき正解となる結果が必要であり，そのためにはやはり膨大な回数のシミュレーションを実行する必要がある．また，提案法によりシミュレーション回数を削減したとしても，ネットワークが大規模化した場合は個々のシミュレーシ

ョンの計算時間が膨大なものとなる．そのため，本研究では，ネットワーク中のノードの影響度計算が，一定の条件下の当該ネットワークにおいて，そのノードからの可到達ノード集合を求めることと同じであることに着目し，個々のノードが可到達ノード集合（初期値は隣接ノード集合）の情報を単位時間ごとに隣接ノードに相互伝播し，その情報を更新する計算モデルを検討し，さらに，そのノードごとの処理を並列分散環境で効率よく計算するアルゴリズムを検討した．

４．研究成果

（1）リサンプリング法に基づいた予測シミュレーション法に関する成果

予測シミュレーションの予測精度保証に関する理論的枠組みに関しては，リサンプリング法の１つである leave-N-out 法を基礎とし，複数回のシミュレーション結果からサンプリングした一定数の結果に基づきその時点でのシミュレーション結果の精度を推定する手法を提案した．社会ネットワーク中のノードがもつ影響度の推定を対象とした実験では，これまで安定した精度を得るために盲目的に1万～10万回実行していたシミュレーションを高影響度ノードの抽出という目的の下では 100 回程度に減らせる可能性を示すことができた．

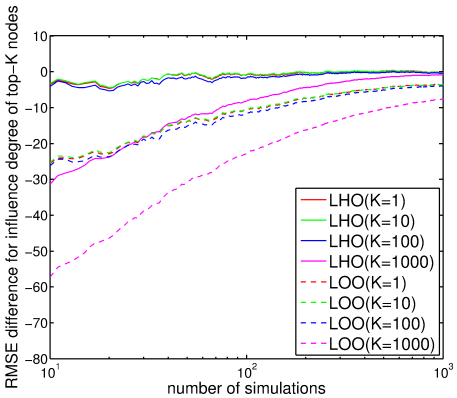


図１：影響度上位 K ノードに対する提案法による推定近似誤差の評価

評価実験の結果の一部を図１に示す．グラフの横軸はシミュレーション回数，縦軸は真の近似誤差に対する提案手法による推定近似誤差の RMSE（二乗平均平方根誤差）を表している．LHO は提案手法のうち各時点のシミュレーション結果の半分をサンプリングするものであり，LOO はすべてのシミュレーション結果を１度ずつサンプリングする交差検証法を実行するものである．この結果は，LHO 法を用いることで，高影響度ノード上位 K =100 程度に関しては，1,000 回のシミュレーションで得られる精度を数十回のシミュレーションで精度よく近似できていることを示すものである．

（2）リサンプリング法に基づいた信頼度つきギャップ分析を用いた高中心性ノードの同定に関する成果

前述のリサンプリング法に基づいた予測シミュレーション法の方法論を社会ネットワークにおけるノードの重要性指標であるノード中心性の計算問題に応用し，中心性の高い重要ノード群を一定の精度保証付きで推定するためのギャップ分析法を提案した．このギャップ分析法では，ネットワーク中の全ノード間に対して実行すべき最短パス長の計算などの基本計算を，サンプリングされた一部のノード間のみで実行し，その結果に基づき中心性指標の大小関係が一定の精度で保証されるノード群を同定する（ノード群間にギャップが存在すると考える）．これは，大規模ネットワークでは予測シミュレーションと同様に基本計算を反復することにより膨大な計算時間を要する中心性指標の計算を回避し，重要ノード群の特定を可能にするものであり，同時に，提案する予測シミュレーションの枠組みが汎用的であり， 反復計算を多用するような類似する実問題に広く応用可能であることを示すものである．

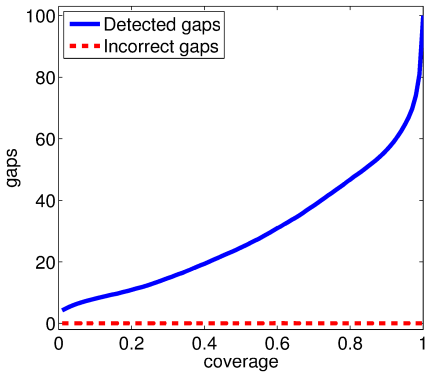


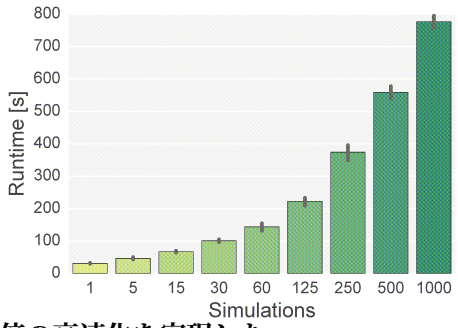
図 2：Twitter ネットワークに対する媒介中心性の計算におけるギャップ同定精度

評価実験の結果の一部を図 2 に示す．グラフの横軸は媒介中心性の計算のために利用したノード数の全ノード数に対する割合であり，縦軸は同定したギャップ数を表している．この結果から，提案手法では全体の約 4 割程度のノードを利用して 20%ほどのギャップをほぼ誤判定することなく同定できていることがわかる．実際には，中心性指標もその値が高いノード群が特定できればよいため，20%ほどのギャップでも実用的であると言える．

（3）社会ネットワークにける影響最大化問題を解く並列分散アルゴリズムに関する成果

前述のようなノードごとの処理を中心とした枠組みで，情報拡散シミュレーションを複数計算機上で並列分散実行するアルゴリズムを開発した．提案アルゴリズムでは，テ

スト用ネットワークに対して 1,000 回の情報拡散シミュレーションを逐次実行する場合に比べて 4 台の計算機での並列分散実行で約



40 倍の高速化を実現した.

図 3：影響度計算シミュレーションの並列分散実行時におけるシミュレーション数と実行時間の変化

シミュレーション回数に対する実行時間の変化を示すグラフを図 3 に示す．この結果から，通常はシミュレーション数に正比例して増加することが見込まれる計算時間の増加率よりがはるかに低いことが分かる．

以上のように，本研究を通して，リサンプリング法に基づく予測シミュレーション法とその応用，および並列分散実行に関する多くの知見を得ることができた．ここで得られた知見は，今後，ますます大規模化するであろうインターネット上の社会ネットワークの分析を効率よく進める上で，重要な要素技術となり得るものである．今後，より多様なデータを対象にさらに具体的な知識発見への応用を進めていく予定である．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計　8　件）

① 佐々木 亮輔、豊田 哲也、大原 剛三、社会ネットワークにおける影響最大化問題を解く並列分散アルゴリズムの提案、第 9 回データ工学と情報マネジメントに関するフォーラム（DEIM2017）、2017 年 3 月 8 日、高山グリーンホテル（岐阜県・高山市）

② Kouzou Ohara, Kazumi Saito, Masahiro Kimura and Hiroshi Motoda, Resampling based Gap Analysis for Detecting Nodes with High Centrality on Large Social Network, The Nineteenth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2015), 2015 年 5 月 21 日, Ho Chi Min City (Vietnam)

③ Kouzou Ohara, Kazumi Saito, Masahiro Kimura and Hiroshi Motoda, Resampling based Framework for Estimating Node Centrality of Large Social Network, The Seventeenth International Conference on Discovery Science (DS2014), 2014 年 10 月 10 日, Bled (Slovenia)

④ 大原 剛三、斉藤 和巳、木村 昌弘、元田 浩、社会ネットワーク上の強影響度ノード同定のためのリサンプリングに基づく予測シミュレーション法の提案、第 28 回人工知能学会全国大会（JSAI2014）、2014 年 5 月 15 日、ひめぎんホール別館（愛媛県・松山市）

６．研究組織

(1)研究代表者

大原　剛三　（OHARA, Kouzou）

青山学院大学・理工学部・准教授

研究者番号：30294127

(2)研究分担者

斉藤　和巳　（SAITO, Kazumi）

静岡県立大学・経営情報学部・教授

研究者番号： 80379544

木村　昌弘　（KIMURA, Masahiro）

龍谷大学・理工学部・教授

研究者番号：10396153