4O1-OS-17a-2in2

クラウドソーシングのみによる因果関係発見の試み

米良俊輝 *1 Toshiki Mera　若宮翔子 *2 Shoko Wakamiya　荒牧英治 *2 Eiji Aramaki　森嶋厚行 *3 Atsuyuki Morishima

*1筑波大学大学院 図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba

*2奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

*3筑波大学 知的コミュニティ基盤研究センター
Research Center for Knowledge Communities, Faculty of Library, Information and Media Science, University of Tsukuba

A task of forming and verifying a hypothesis on human condition, such as “stress is the cause of depression” is costly because it is not easy to perform the task without experts. In this paper, we propose a method to use microtask crowdsourcing to obtain hypotheses and verify causal hypotheses. First, we crowdsource enumerating hypotheses, and concurrently try to rank the enumerated ones in order of their probability. Next, regarding hypotheses which seem to be probable, we try to infer causality to consider a temporal bias to have the crowd perform a task of trying out contents of the hypotheses and reporting the results. We conduct an experiment using an actual crowdsourcing service, and evaluate the results to examine how much scientific evidence they have.

1. はじめに

ある事実と別の事実との間の因果関係を発見することは，科学や医療の発展，政策の決定などにおいて非常に重要である．安価に大量のデータを収集することが容易になりつつある近年では，母集団とほぼ等しいサイズのデータを解析することで因果関係を発見する試みも行われている．しかし，そのように大量のデータを安価にかつ簡単に入手可能な領域は必ずしも多くない．

大量のデータを容易には入手できない領域では，データを集めつつ因果関係を発見することが必要となるが，そのためにこれまでは，専門家が仮説を立てて実験を行い検証するというプロセスが実施されてきた．専門家がこうした作業を実施する際には，直接的な雇用のコストのほか，問題ごとにチームを組織するなどの面でもコストがかかる．この専門家による作業を非専門家からなる群衆によって代替し，仮説を入手・検証することが可能となれば，調査研究分野の進展に貢献することが期待できる．

そこで本研究では，非専門家へのクラウドソーシングによって，仮説を入手しそれらを検証するという因果関係の発見に必要なプロセスを実現することに取り組む．また，クラウドソーシングのみによって因果関係を発見する上で，それがどのレベルの科学的根拠をもって実現可能なものであるかを明らかにすることを目的とする．そのために，(1) 検証すべき仮説の選別 (2) 時間的前後関係の検証 を 2 段階のタスクを用いて実施することで，クラウドソーシングと計算機処理のみで因果関係の推論を行う手法を提案する．

また提案手法によって，クラウドソーシングで得た情報から因果関係の推論をどのレベルの科学的根拠をもって実施できるのか，実験を実施し結果の評価について報告する．

本稿の貢献は次の通りである．

1. クラウドソーシングと計算機処理のみで，仮説形成から因果推論を含めた検証までの実施を試みた初めての研究である．

2. 実際のクラウドソーシングによる実験を実施し，その結果を用いて検討を行った．

連絡先: 米良 俊輝, 筑波大学, 〒 305-8550 茨城県つくば市春日 1-2，029-859-1559，toshiki.mera.2015b@mlab.info

2. 関連研究

クラウドソーシングを用いて何らかの仮説やアイデアを入手し問題解決に役立てることに関しては，その可能性・展望 [1][2] や，具体的な事例について [3][4][5] 研究が行われてきた．入手した仮説の検証をクラウドソーシングで行う研究についても，これまで医学研究の分野で実施されている [6]．いずれの研究も，群衆から得た情報を活用しているものの，それらの検証までは実施されていないか，あるいは専門家により実施されている．本研究では，群衆から仮説を入手し，さらにそれらの検証も群衆の力で実施する．また本研究は，3.2 節に示すように，仮説検証をこれまでの研究より 1 段階高いレベルで実現している．

本研究で扱う問題は，仮説という評価が未確定のものに対して検証を行いつつ評価の高そうなものを優先的に選択しようとしている点で，探索と活用の両立を考える多腕バンディット問題と関連がある．クラウドソーシングのタスクにおいて多腕バンディットと関連した問題を解くモデルを提案した研究 [8] も存在する．これに対し本研究では検証すべき仮説を 1 回のタスクで複数個選択できるなどの点で，単純に多腕バンディット問題のアルゴリズムを利用することはできない．

3. 非専門家へのクラウドソーシングによる因果推論とその限界

本節ではクラウドソーシングを用いた因果推論が可能な条件とその限界について述べる．本研究では非専門家を対象としたクラウドソーシングを扱うが，ここでの非専門家とは，文献調査などは行わずに自身の経験に基づいてのみ回答する人々とする．本節では，このような人々にクラウドソースすることで因果関係を発見する手法がどのような範囲の問題に対して適用可能であるか，またどのようなレベルで実現可能なのかについて説明する．

表 1: エビデンスレベルとクラウドソーシングによって実現可能なレベル

レベル	内容	従来	提案
1	ランダム化比較試験		
2	非ランダム化比較試験		
3	分析疫学的研究 (コホート研究)		○
4	分析疫学的研究 (症例対照研究)	○	○
5	前後比較試験	○	○
6	症例報告 (記述的研究)	○	○
7	専門家個人の意見		

3.1 適用可能な問題の条件

提案手法を適用可能な問題の条件は次の 2 つである．1 つ目は，入手する仮説が回答者本人が認識可能な状態に関するものであることで，本人が認識可能な状態とは，例えば環境・行動・経験・体質などを指し，遺伝子異常のような本人が認識できない状態に関する仮説は含まれない．2 つ目は，群衆の中に仮説が示す状態の該当者が存在することである．

3.2 クラウドソーシングによる統計調査の限界と可能性

クラウドソーシングの調査結果を用いて因果関係が本当に存在するかどうかを証明することは現実的には困難である．本研究で扱うクラウドソーシングを用いた調査は，人間を対象とすることやインターネットを介した調査であることから，実験群と対照群それぞれにワーカをランダムに割り振る無作為割り当てを行えないため，実験研究ではなく調査観察研究となる．調査観察研究では，原因となる変数を研究者が正確に操作できないことから様々な偏りが生じるため，因果関係を証明することは出来ない．そこで，考えられる偏りによる影響を取り除き，どの程度の科学的根拠をもって因果関係を推論することができるかについて考える．

調査結果の偏りをどの程度取り除くことができるか (因果関係を示す力の強さ) に応じて研究デザインの序列を設けた基準としてエビデンスレベルがある．主に医学研究分野では，可能な限り良いエビデンスを意思決定で慎重に用いる考え方が重視されており [9]，このエビデンスレベルがしばしば利用される．

本研究では人間の状態に関する仮説を扱うため，エビデンスレベルを参考として用いて既存手法と提案手法の比較を行った．表 1 に示したエビデンスレベルは，一般的に上のものほど因果関係を示す力が強いことを表す．従来の調査においてクラウドソーシングによって実現できる範囲は，現在と過去に起こった事象を集め分析することまでで，表 1 のレベル 4 から 6 に該当するものであった．一方，本研究ではこれに加えて未来の事象を追跡して分析することでレベル 3 に該当する調査を実現する．

4. 提案手法

本節では，因果関係に関する仮説を入手するために用いる 2 フェーズの手法について説明する．提案手法の概要を図 1 に示す．実際に仮説を入手し選別するフェーズ 1，選別した仮説を検証するフェーズ 2 について説明する．

4.1 フェーズ 1 ：検証すべき仮説を選別

フェーズ 1 の入力と出力は次の通りである．

入力： 回答者の状態が A か $\neg A$ かを判定する質問，仮説を入手するための質問

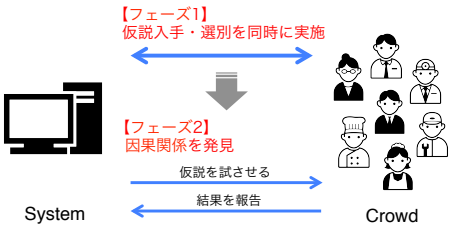

図 1: 提案手法の概要

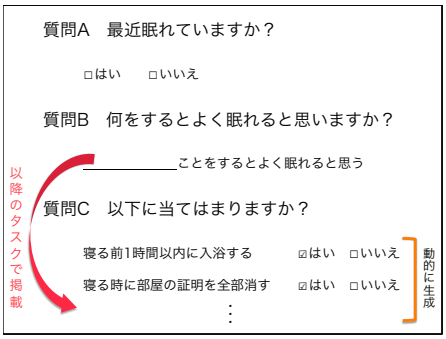

図 2: 仮説入手タスクで用いる質問文の例

出力： A の原因と考えられる仮説のランキング $[h_m, h_n, ...]$

フェーズ 1 では，実際に群衆から仮説を入手し，入手した中から検証するべき仮説を選択するためのタスクを実施する．本節では，このタスクへの回答をもとに仮説の確からしさをどのように定義し，仮説を選択していくかについて述べる．

4.1.1 仮説入手タスク

仮説入手タスクでは，群衆からの仮説入手とそれらの選別を同時に実施する．タスクは図 2 のようなアンケート形式で，次に示す三つの質問群からなる．

質問 A）アウトカム質問： 回答者の状態が A か $\neg A$ かを判定するための質問．

質問 B）仮説プール投入質問： 回答者から仮説 h_i を得る質問．

質問 C）仮説検証質問： 質問 B で以前に得た仮説を掲載することで，その仮説に当てはまるかどうかを尋ねる質問．

質問 A および質問 B の内容は固定で，質問 C の内容のみ動的に生成される．

4.1.2 仮説の確からしさ

4.1.1 節で述べたタスクの回答をもとに，仮説の確からしさを数値化するための指標について定義する．ある仮説 h_i の確からしさを表す値を $P(h_i)$ とする．これは，表 2 のように，質問 A で判定した回答者の状態と，質問 C への回答数とのクロス集計表 T の値から計算する．すなわち，この P は，クロス集計表 T の値を引数とする関数 f として，$P(h_i) \equiv f(a, b, c, d)$ と定義する．この $P(h_i)$ の値が大きいほど確からしい仮説であるとみなす．f の計算に用いる具体的な指標は，扱う問題に応じて適当なものを用いる．

4.1.3 仮説選択の目標

4.1.1 節の質問 C について，ある時点までに入手したすべての仮説を検証する必要は必ずしもないため，質問 C には検証すべき仮説を一定数選択して掲載する．本手法では，仮説選択の目標を次のように設定する．

表 2: $P(h_i)$ の計算に用いる集計表 T の例

		仮説 h_i	
		該当する	該当しない
状態	A	a (人)	b (人)
	$\neg A$	c (人)	d (人)

表 3: 相関計算に用いる集計表の例

		仮説 h_i	
		該当する	該当しない
仮説 h_j	該当する	r (人)	s (人)
	該当しない	t (人)	u (人)

Algorithm 1 仮説選択のアルゴリズム

Input: m, H_t
Output: H'_t
1: $H'_t \leftarrow \phi$
2: **while** $(|H'_t| < m) \wedge (H_t \neq \phi)$ **do**
3: $h_i \leftarrow getByWRandom(H_t)$
4: delete h_i from H_t
5: **if** $checkCorrelate(h_i, H'_t)$ is true **then**
6: add h_i to H'_t
7: **end if**
8: **end while**

目標 1 確からしい可能性の高い仮説を選択: 最終的に有用な仮説として上位にランキングされる可能性が高いものを優先して選択したい．

目標 2 内容が本質的に同じ仮説を同時に選択しない: 本質的には同じ内容の仮説ばかりを一つのタスクで掲載すると，内容が本質的に異なる他の仮説が選択される機会を奪うことになるため，内容が本質的に同じ仮説を除外して選択したい．

4.1.4 仮説選択アルゴリズム

本節では 4.1.3 節で述べた 2 つを目標に，各タスクに掲載する一定数の仮説を選択する手法を提案する．具体的には，タスクを発行する際に，Algorithm 1 を呼び出す．これは，ある時点 t で入手済みである仮説の集合 H_t の要素のうち，次のタスクで質問 C として掲載するための m ($\in \mathbb{N}$) 個を選択するアルゴリズムである．選択済みの仮説の集合は H'_t と定義する．

目標 1 実現のための手法: ある時点での $P(h_i)$ の値に基づく重み付きランダムで仮説を選択する．この重み付きランダムで，ある仮説 h_i が選択される確率 $Q(h_i)$ は次で定義される．

$$Q(h_i) = \frac{P(h_i)}{\sum_{h_j \in H_t} P(h_j)}$$

これはある時点で確からしい仮説は最終的にも確からしい仮説である可能性が高い，というヒューリスティクスに基づく手法であり，Algorithm 1 中の getByWRandom 関数はこの手法を実装している．

目標 2 実現のための手法: H_t から H'_t へ追加する候補として h_i が選択されたとき，h_i と h_j に相関があるような h_j がその時点で H'_t に存在した場合は，h_i を H'_t には追加しない．具体的には，質問 C の回答から作成した表 3 のようなクロス集計表の値から相関を計算する．これは，内容が重複する仮説は回答に強い相関が出るというヒューリスティクスに基づく手法である．Algorithm 1 中の checkCorrelate 関数では，h_i と H'_t を引数として受け取り，h_i と全ての $h_j \in H'_t$ との間で相関を計算する．そして相関のある組み合わせが 1 つでも存在すれば false を，1 つも存在しない場合には true を返す．

4.2 フェーズ 2 ：選別した仮説を検証

フェーズ 2 の入力と出力は次の通りである．

入力 : 検証を行う仮説の集合 $\{h_1, h_2, ..., h_k\}$
出力 : A の原因だと示された仮説の集合 $\{h_x, h_y, ...\}$

フェーズ 1 では，ある問題と仮説との間の相関を単純に調べただけであるため，擬似相関や因果の向きが逆である場合などを考慮できていない．そこでフェーズ 2 では，仮説の内容を群衆に試してもらい，現在から未来にかけてのデータを分析することで，より高いレベルで因果関係を推論する手法の実現を目指す．

4.2.1 仮説検証タスク

フェーズ 1 で出力したランキングで上位の仮説について，その内容を群衆に試して報告してもらう過程で 2 段階のタスクを実施する．1 段階目のタスクでは仮説 h_i を試す以前のワーカの状態 S_{h_i1} を確認し，2 段階目では試した後の状態 S_{h_i2} を確認する．$S_{h_i2} - S_{h_i1}$ が有意であるか確認することで，表 1 のレベル 3 で因果関係を検証することを試みる．

5. 評価実験

本節では，4. 節で提案した手法の評価を行うため，各フェーズに対する実験結果について述べる．

5.1 実験概要

評価実験では，Crowd4U[*1] 上でタスクを作成し，Yahoo!クラウドソーシング [*2] で参加者を募ってタスクを依頼した．

今回設定したタスクは，多くの人が関係し，かつ状態を定量的に評価する指標が存在する「睡眠の質」の問題について，「ぐっすり眠れるような仮説」を入手し検証するものである．回答者の状態の判定にはピッツバーグ睡眠質問表 (PSQI) [10] を用い，仮説の確からしさを示す値には医学分野の関連研究 [3][6] でも用いられているオッズ比を使用した．

5.2 仮説入手タスクと仮説検証タスクのパラメータ

4.1.1 節の仮説入手タスクを用いた手法を実装する際に設定したパラメータを次に示す．

1 タスクで入手する仮説数 : タスク内の質問 B を「自分は何をすればよく眠れるか」と「何をすればよく眠れると思うか」の 2 問とし，1 つまたは 2 つの仮説を回答するものとした．

仮説間の相関を求める検定手法 : 4.1.4 節の相関があるかどうかの判定には，Fisher の正確確率検定を用いた．

仮説入手タスクに掲載する仮説数 : 質問 C として掲載する仮説数は 10，すなわち 4.1.4 節の $m = 10$ とした．

新規入手仮説へ割り当てる $P(h_i)$ の値 : オッズ比を使用すると，入手直後の仮説の $P(h_i)$ の値が非常に小さくなる．本実験では入手した仮説に対して，回答数が 10 未満の場合にはその時点で存在する $P(h_i)$ の最大値を割り当てた．

検証する仮説数 : フェーズ 1 で入手・選別を行った仮説のうち，フェーズ 2 で検証する数は 10 とした．

仮説検証タスクの間隔 : 仮説検証タスクにおいて，1 段階目を実施してから 2 段階目のタスクを依頼するまでの期間は 1 週間とした．

*1 http://crowd4u.org/
*2 http://crowdsourcing.yahoo.co.jp

表 4: タスク結果の評価 (下線：有意水準 5 %未満)

専門家判断	仮説	回答数	p-value1	p-value2
効果あり	1	21	1.0000	1.0000
	2	55	0.3498	<u>0.0008</u>
	3	28	1.0000	0.1440
	4	40	0.3332	0.6339
	5	26	0.3304	1.0000
効果なし	6	22	0.2536	0.2281
	7	18	1.0000	0.6447
	8	50	0.3725	<u>0.0201</u>
	9	60	<u>0.0217</u>	0.1534
	10	22	1.0000	0.3512

5.3 仮説検証タスクで提示する仮説の選択

タスクで提示する 10 個の仮説を選ぶ上で，(1) まず医療従事者がランキング上位 100 個のうち倫理的に問題のある仮説を除外し，(2) さらにそれぞれに「効果がありそう」または「効果がなさそう」のラベル付けを行った．(3) そして実際に提示する際には，内容が重複しないように，2 種類のラベルから 5 個ずつ計 10 個の仮説を人手で抽出した．その際に，回答基準のばらつきを抑えるため，できるだけ具体的な内容のものを選び,「寝る直前」「寝る前に」のような曖昧な表現の箇所にそれぞれ「 (5 分以内) 」「 (1 時間以内) 」といった語句を付加した．なお，ここでの (1) は必ずしも医療の専門家でなければできない作業ではなく，(2) は本稿における手法評価のための作業であり，手法そのものには不要である．また，(3) については医療従事者でない著者らが実施した．

5.4 実験結果

仮説入手タスクでは，1546 件の回答を得た．続く仮説検証タスクにおいて，1 段階目のタスクでは 686 件，2 段階目のタスクでは，343 件の回答を得た．最終的に得た 343 タスク分の回答結果を分析したものを表 4 に示す．

p-value1： 仮説の内容を週の過半数の日で実施できたかどうかと，PSQI スコア (数値が低いほど睡眠の質が良い) が下がったかどうかで回答をクロス集計する．この結果で Fisher の正確確率検定 (両側検定) を実施した際の p 値．

p-value2： 仮説の内容を週の過半数の日で実施できたかどうかと，閾値 (6 点) をまたいで PSQI スコアが低下したか，すなわち状態が不眠から健康へ変化したか否かをクロス集計した結果に，前述の検定を実施した際の p 値．

p-value1 の値によると，効果なしのラベル付けがされた仮説「寝る直前 (5 分以内) にゆっくりと深呼吸する」のみが有意水準 5%で有意となった．p-value2 の値によると，効果ありのラベル付けがされた仮説「寝る直前 (5 分以内) に室内の照明をすべて消す」と，効果なしのラベル付けがされた仮説「寝る前 (1 時間以内) にリラックスができる静かな音楽を聴く」が有意水準 5%で有意となった．また，専門家の判断により効果ありとされたグループと効果なしとされたグループの間で，いずれの群衆による検定結果にも有意な差は確認できなかった．

5.5 考察

5.4 節の結果から，群衆による仮説検証の結果は専門家による判断とは異なることが考えられ，さらに専門家が効果なしと判断した仮説の中からも，有意に因果関係があると考えられるものが発見された．このことから，非専門家によるクラウドソーシングのみによって専門家では発見できないような仮説を発見できる可能性が示唆された．

6. まとめ・今後の課題

本稿では，クラウドソーシングにおいて従来より高いレベルで因果関係に関する仮説を発見するために，2 フェーズからなる手法を提案し，実際に群衆にタスクを依頼する実験でその手法を評価した．その結果から，専門家では発見できなかったような因果関係に関する仮説を群衆によって発見できる可能性が示唆された．

今後の課題として，既存研究との比較の際に用いたエビデンスレベルはあくまで一般的な研究デザインの評価に段階を付けたもので，個別の研究ごとに条件や変数などの設定次第で評価は入れ変わるため，より詳細な項目や基準でエビデンスの質を評価することが挙げられる．また，実験のランキング結果に同じ内容の仮説や表現が曖昧な仮説が多く存在し，試させる仮説をある程度人手で選別したため，この部分に関して自然言語処理などを活用し群衆と計算機処理のみで実施できるか検討する予定である．さらに，本研究で仮説入手に用いるタスク数は固定値だが，クラウド DB における列挙型クエリの進捗を推定する研究 [7] などの知見を生かすことで，必要なタスク数を動的に判断する手法についても検討する．

謝辞

本論文の一部は JST CREST および科研費基盤研究 (#25240012) の支援による．

参考文献

[1] Swan, M. “Crowdsourced health research studies: an important emerging complement to clinical trials in the public health research ecosystem.” Journal of medical Internet research 14.2 (2012): e46.

[2] Brabham, Daren C. “Crowdsourcing as a model for problem solving: An introduction and cases.” Convergence 14.1 (2008): 75-90.

[3] Bevelander, Kirsten E., et al. “Crowdsourcing novel childhood predictors of adult obesity.” PloS one 9.2 (2014): e87756.

[4] Bongard, Josh C., et al. “Crowdsourcing predictors of behavioral outcomes.” IEEE Transactions on Systems, Man, and Cybernetics: Systems 43.1 (2013): 176-185.

[5] Poetz, M. K., Schreier, M. “The value of crowdsourcing: can users really compete with professionals in generating new product ideas?.” Journal of Product Innovation Management 29.2 (2012): 245-256.

[6] 荒牧英治, et al. “クラウドソーシングによるアレルギー・リスク推定:-仮説形成から実験までの研究を半自動化する試み.” 研究報告自然言語処理 (NL) 2014.22 (2014): 1-6.

[7] Trushkowsky, B., et al. “Crowdsourced enumeration queries.” Data Engineering (ICDE), 2013 IEEE 29th International Conference on. IEEE, 2013.

[8] Abraham, I., et al. “Adaptive Crowdsourcing Algorithms for the Bandit Survey Problem.” COLT. 2013.

[9] 福井次矢, et al. “Minds 診療ガイドライン作成の手引き 2007.” Minds 診療ガイドライン選定部会監修, 医学書院, 東京 (2007).

[10] Buysse, Daniel J., et al. “The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research.” Psychiatry research 28.2 (1989): 193-213.