2H1-1in1

人狼ゲームにおける信頼の分析

Analysis of trust in “Are you a Werewolf?”

園田 亜斗夢　鳥海 不二夫
Atom Sonoda　Fujio Toriumi

東京大学
The University of Tokyo

Artificial Inteligence (AI) has been introduced into the world. However, there is a need to improve the precision of these applications and to establish a mutal trust between AI and human beings. In this paper, we analyze the trust in “Are you a Werewolf?” as a game that specializes in trust and persuasion. Finally, we found that the players can find out each other's real role that is not previously explicit and made their decision based on which player they trust.

1. 序章

現在，我々の生活には続々と人工知能が導入されている．しかしながら，人間とエージェント間のコミュニケーションは未だ十分であるとは言えず，社会からの信用も低い．そこで，人工知能が人間から信頼を得るにはどのようなコミュニケーションが必要かを明らかにする必要がある．

そこで，本論文では，短期的信頼関係構築のメカニズムを明らかにし，説得を可能にするエージェント技術の実現するために，信頼と説得に特化したコミュニケーションゲームである人狼ゲーム [1] に注目する．

人狼ゲームは対話のみを通じて信頼の醸成や意図の推測，説得を行うゲームである．よって，このように限定された環境である人狼ゲームを対象とすることで，信頼の分析が可能となる．本研究では，そのような人狼ゲームを対象として，ゲームプレイの中で短期的な信頼がどのように確立していくのかを明らかにする．その中でも，特にゲーム中で重要な役割を担う役職「占い師」に着目し，プレーヤーの情報や発言の特徴や発言内容と投票行動の関係を時系列的に捉えることで，プレーヤー間での説得の成否や信頼構築の推定を行う．これにより，高度なコミュニケーション能力に必要な要素を解析することが可能になり，コミュニケーションに関する研究に貢献できると考えられる．

連絡先: 園田亜斗夢，東京大学工学部システム創成学科，113-8656 東京都文京区本郷 7-3-1 工学部 8 号館 526，TEL: 03-5841-6991，E-mail: sonoda@crimson.q.t.u-tokyo.ac.jp

2. 関連研究

コミュニケーションと信頼・説得に関する研究は，数々の研究がなされている．E コマースの分野で広く使われている協調フィルタリングなどの推薦技術では，ユーザの属性や情報の提供方法により情報の受け入れやすさが異なることが確認されている [2]．しかしながら，基本的にユーザと一対一で説得する場合のみで，競合が存在しどちらを信頼するかといった状況は考えられていない．

また，人狼に関する研究も，数々の研究がなされている．人狼をプレイするエージェントの分析から，人狼の推定には占い情報が重要で，人狼推定の精度が向上することが村人陣営の勝利につながることが確かめられている [3]．また，人間がプレイしたログデータの分析から人狼ゲームの基本的性質の分析がなされている [4]．しかしながら，これらの分析では，どのようなコミュニケーションが説得に影響を与えるのかという分析話されていない．

そこで，本研究では，嘘が含まれる状況で議論が進み，またプレーヤーの人数が多いという点で複雑性の高いゲームである人狼ゲームにおいて，どのようなプレーヤーが信頼されるのか分析する．

3. 信頼獲得の特徴の分析

本章では，各プレーヤーの発言がどのように信頼獲得に影響を与えているのかについて分析する．人狼ゲームでは説得や議論を通してゲームが進められ

るため，短期的信頼を獲得しようとする場面は多く存在し，他のプレーヤーの役職を推定する場面では説得が重要になる．特に，占い情報は村人陣営にとって最も重要な情報のうちの一つである．そこで，特徴別の信頼の獲得について分析する．

3.1 分析対象ログデータ

人狼をプレイする方法としては， 市販のカードなどを使って配役を決め会話を使って行う対面型と，WEB 上のアプリケーションを使って行う電子掲示板型 (BBS タイプ) が存在する．その中でも，日本で多く遊ばれているサービスのうちの一つが人狼 BBS である．

人狼 BBS ではこれまでに 5 千回以上のゲームが行われており，本研究では，このログデータを用いて解析を行った．

3.2 信頼度の定義

本研究で採用した定義では，ユーザとシステム間のコミュニケーションや，システムの動作原理に対して信用を置くことを信頼と定義している．これを具体的に人狼ゲームでの行動に当てはめて考えてみると，発言が信用され，その発言により他のプレーヤーの投票行動などを変えることができているかどうかが，信頼の獲得の指標となる．特に，人狼ゲームの性質上，占い師の判定結果は議論の流れを左右し，勝敗に影響を与えることが確かめられている [3] ことから，本研究では占い師の発言の信用に注目する．そこで，これ以降は最初の投票が行われるまで (2 日目の終わりまで) の占い師の発言と，村人陣営からの投票にのみ着目し，信頼の定義として以下を用いる．

信頼の場合は下の 3 つの条件を全て満たす場合，信頼とする．また，この条件を満たさない場合，疑いとする．

1. 占い師プレーヤー自身の得票が 0
2. 占い師の初日判定が人間判定の場合，判定先プレーヤーの村人陣営からの得票が 0
3. 占い師の初日判定が人狼判定の場合，判定先プレーヤーの村人陣営からの得票が半分以上

特に怪しまれない限り，初日に占い師 CO プレーヤーが村人陣営から投票されることは少ない．これは占い師の情報がそれだけ村人陣営に重要だからであり，もし 1 票でも投票されているとしたら疑われていると言える．また，初日は情報が少ないため占い師に人間判定されたプレーヤーは人間である確率が他のプレーヤーに比べて高いと言えるため，占い師の発言が信用されていれば投票されることはない．また，人狼判定されている場合は人狼である確率が他のプレーヤーに比べて高いと言えるため，占い師の発言が信用されていれば投票を集めるはずである．これらの理由から上記の定義を定めた．これ以降，これらの定義を用いて分析を進める．分析対象のゲームは，自分で初回の投票より先に CO している占い師 CO が 2 プレーヤーのものを対象とした [*1]．この定義を満たすゲームは，586 ゲームだった．その内訳は，表 1 の通りである．

表 1: 分析対象のゲームの内訳

	ゲーム数
信頼と疑いに分類されたもの	184
両方信頼と分類されたもの	19
両方疑いと分類されたもの	383

表 2: 実際の役職と信頼度の関係

実際の役職	信頼	疑い
占い師	112	72
狂人	63	83
人狼	9	29

3.3 実際の役職と信頼度の関連性

信頼される占い師 CO プレーヤーが本物であるか偽物であるかは勝利に重要な影響を与える．そこで，本物か狂人か人狼かに分けて分析する．信頼と疑いに別れた 184 ゲームについて，表 2 に，実際の役職についてそれぞれ前節で定義した信頼と疑いに分けられたゲーム数の内訳を示す．

実際の役職が占い師であるか人狼陣営であるかは信頼される確率について有意水準 1 %で有意な差が認められた．2 日目までは襲撃も追放も行われておらず，霊能結果もないため，本物の占い師と偽物の占い師を見極める客観的な情報は 3 日目以降より少なく，限られている．それにもかかわらず，本物の方が信頼を獲得しやすいことがわかる．

以上より，初回投票までという限られた発言数の中でも，人間は実際の役職を何らかの方法で見抜き，信頼するか疑うかを判断している可能性が高いことが明らかとなった．

*1 適切に情報が取得できた 2011 ゲーム中，1083 ゲームが占い師 CO が 2 プレーヤーで，そのうち 784 ゲームが自分で CO しており，586 ゲームが初日に CO しているゲーム．

表 3: LDA によってトピック分類された単語の抜粋

ラベル	トピックを構成する単語
初日 CO	CO，占い師，霊，狼，初日
初日以降	CO，狂人，占い，タイミング，夜明け
推定	狼，灰，印象，理由，気
判定	狼，白，黒，霊，占い
その他	嬢，疲れ，悪意，ペタ，パン

4. 言語情報による信頼獲得の分析

本章では，より発言の中身に注目し，言語情報から信頼の獲得について分析を進める．

4.1 発言トピックと信頼度の関係

ここでは，統計的潜在意味解析を可能にする Latent Dirichlet Allocation(LDA) を用いて，信頼と分類されたプレーヤーとそうでないプレーヤーについて，発言に含まれる潜在的意味 (トピック) と信頼度の関連性について分析する．LDA[5] は，自然言語処理 (natural language process- ing，NLP) で用いられるトピックモデルの代表的手法である．トピックとは，単語の 共起性に基づき決められる意味カテゴリのことである．トピックに関しては，トピックを構成する上位の単語を人間が見て，トピックラベルを付与した．本研究では推定精度と分析のしやすさからトピック数は 5 とした．

表 3 に，LDA によってトピック分析した時のトピックを構成する単語の抜粋を示す．トピック「初日 CO」には，「占い師」，「CO」といった単語以外に「初日」など初日 CO に関する単語と考えられるものが含まれている．これは，２日目までに占い師 CO しているゲームの発話に限って分析しているため，CO のタイミングを初日にするような発話が多く取れているのだと考えられる．これは，信頼されたプレーヤーの全発言の 25 %以上を占める．人狼ゲームの占い師の発話ではやはり CO に関する内容が多いことが確認できる．占い結果は最初の追放までの村人陣営の唯一の客観情報であり，その重要性により，発話も増えるのだと考えられる．トピック「推定」は，「印象」「気」という能力ではない理由で「白」，「狼」など陣営を推定しているのだと考えられる．これは，信頼されたプレーヤーの全発言の 16 %以上を占める．

これらのトピックに関して，信頼と疑いで発言に含まれるトピックの割合の平均に有意差があるかどうか分析した．ここではトピック「初日 CO」と「推定」が 1 %有意で差が見られた．トピック「初日 CO」は信頼プレーヤーに多く見られ，トピック「推定」は疑いプレーヤーに多く見られた．他のトピックには有意差が見られなかった．このことから，初回投票までに重要な戦略や情報となる初日 CO に関する発言が多いほど信頼されていると言うことになる．また，「印象」などの根拠があいまいな推定に関する発言が多いほど信頼されにくいことがわかる．これは，占い師であれば占い結果という根拠のある推理ができるはずなのに，印象に頼った発言をするのが怪しいと思われているのだと推測される．本章では LDA を用いてトピックの集合としてモデル化した．その上で，信頼の獲得別に分析したところ，説得したい内容を根拠を持って話していた方が信頼を獲得しやすいことがわかった．また，感情に関する発言についても有意差が確認された．しかしながら，信頼と疑いの間の差は大きくなく，その差にどれほど意味があるのか確かではない．また，信頼と疑いの間の差が認められないトピックも多かった．

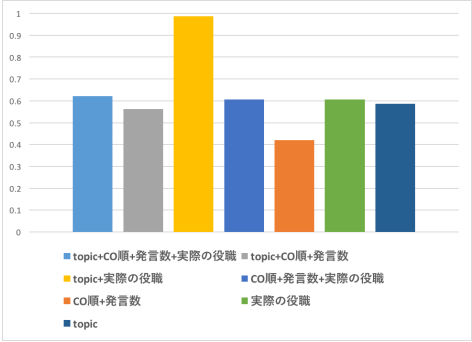

図 1: 特徴量別の正答率

5. 信頼の予測

本章では，人狼ゲームにおいて占い師の CO と判定の信頼の獲得について，それぞれの特徴量と信頼度の関係をる信頼されるプレーヤーの予測手法を提案し，その予測精度の評価を行う．この予測を通じて，人間がどのように信頼と疑いを分けているのかの解明を目指す．

5.1 信頼度推定の評価

図 1 に，利用した特徴量ごとの正答率を示す．この図から，トピックと実際の役職を用いた場合が正答率 0.96 と際立ってよく予測できていることがわかる。また，CO 順と発言数のみの場合は正答率が 0.5 以下とランダムの場合より低くなっていることがわかる．このことから，CO 順と発言数のような

表 4: 役職，トピックを特徴量に用いた場合のロジスティック回帰係数

特徴量	係数
初日 CO	-4.22
初日以降 CO	-0.65
推定	8.36
判定	-5.42
その他	-0.76
役職	-8.72

発言に関する表面的な情報は信頼の判断に影響を与えていないことがわかる．

次に，トピックのみと実際の役職のみのそれぞれの場合について注目すると，実際の役職のみの方がよく予測できている．このことから非公開情報である実際の役職の影響が大きいことがわかる．表 4 に，トピックと実際の役職を用いた場合のロジスティック回帰分析の係数を示す．トピックについては係数が正で大きいトピックの割合が多いほど信頼され，係数が負のトピックが多いほど疑いとされ，また，役職は占い師だと信頼され，それ以外だと疑いとされることを意味する．これからも，実際の役職の影響が最も大きいことがわかる．また，トピックには影響の大きいものと小さいものがあることもわかる．「初日 CO」と「推定」以外に,「判定」の影響も大きく，能力の結果という根拠のある発言を表す「判定」が含まれるほど信頼されるということがわかった．

以上より，信頼の判断には非公開情報である実際の役職の影響が大きいことがわかった．また，それに加えて，トピックも影響を与えていることがわかった．しかし，これらの個々の情報を利用した場合では 0.6 程度の正答率しかなく，トピックと実際の役職を同時に使って推定を行うと 0.96 という正答率になった．つまり，人間は第 4 章で分析した発言の意味内容だけでは信頼を判断しておらず，トピックに表れないような発言から非公開情報の実際の役職を推定し, それに基づいて信頼の判断をしていると考えられる．

6. 結論

本研究では，人狼ゲームにおける短期的な信頼がどのように確立していくのかを明らかにすることを目的として，主に占い師の発言について注目し分析した．信頼と説得に特化したコミュニケーションゲームである人狼ゲームにおいて，短期的な信頼獲得という現象が存在することは確認できた．しかし，発言量や順番などの表面的な情報や LDA で分析できるような簡単な言語情報からは，信頼の獲得について必要な要素を分析することは困難であることがわかった．一方，実際の役職と発言トピックを合わせると信頼と疑いを分類できるということから，実際の役職と発言トピックの論理的な繋がりの分析を進めることで，信頼する占い師を決定するメカニズムが明らかになると期待される．一方で，公開されている情報のみによる推定は精度が低かったことから，信頼の獲得は発言量や順番などの表面的な情報や LDA で分析できるような簡単な言語情報からは明らかにできない複雑システムとしての分析が必要であると考えられる．また，信頼する側の問題としては，嘘の発言を信じることは問題がある．つまり，人工知能の発展に関連して，説得できるシステムの開発も重要だが，簡単に騙されないシステムの開発も重要であるため，信頼の獲得の逆の問題として嘘を見抜くという分析を行うことを今後の課題として挙げる．

参考文献

[1] 篠田孝祐, 鳥海不二夫, 稲葉通将, 大澤博隆, 片上大輔. Fan-14-016 人工知能標準問題としての人狼ゲームの提案 (os: 人狼知能研究). インテリジェントシステム・シンポジウム講演論文集, Vol. 2014, No. 24, pp. 74–77, 2014.

[2] 小柴, 相原健郎, 小田朋宏, 星孝哲, 松原伸人, 森純一郎, 武田英明ほか. 説得性に基づく情報推薦手法の提案: 送り手の属性に着目したモデルと検証. 情報処理学会論文誌, Vol. 51, No. 8, pp. 1452–1468, 2010.

[3] 梶原健吾, 鳥海不二夫, 稲葉通将, 大澤博隆ほか. 人狼知能大会における統計分析と svm を用いた人狼推定を行うエージェントの設計. 2016 年度人工知能学会全国大会 (第 30 回) 論文集, 2016.

[4] 稲葉通将, 鳥海不二夫, 高橋健一ほか. 人狼ゲームデータの統計的分析. ゲームプログラミングワークショップ 2012 論文集, Vol. 2012, No. 6, pp. 144–147, 2012.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022, 2003.