情報処理学会論文誌 Vol.61 No.10 1639–1646 (Oct. 2020)

メタデータを利用したソーシャルメディア内グループのネットリスク検知

西口 真央[1,a)] 鳥海 不二夫[1,b)] 高野 雅典[1,2,c)]

受付日 2019年11月6日, 採録日 2020年7月7日

概要：近年，オンライン空間上において，主に未成年者を対象としたいじめや誘い出しのような犯罪が起こる可能性，いわゆるネットリスクを抑制することが重要な課題となっている．従来の解決アプローチの1つとしては，主に会話コーパスを入力とした，ネットリスクの高いメッセージやユーザを自動検知する取り組みがあげられる．しかしながら，実社会では会話コーパスの利用自体が困難なケースも存在し，かつ，近年は複数人で交流可能なメディアにおけるリスクが顕著に高まっている．本研究では，会話コーパスを用いずに，メタデータのみを利用してネットリスクの高いグループを識別するモデルを構築する．実データを用いた2クラス分類モデル構築実験の結果，Macro-F1値で0.883と高い精度で高リスクグループが検出可能となった．さらにモデルを解釈することで，特定のネットワーク構造を持つユーザが所属するグループはリスクが高まる，などのいくつかの興味深い知見を得た．

キーワード：ソーシャルネットワーク，ネットリスク検知，いじめ，誘い出し

Detection of High Online-Risk Groups on Social Media Using Action Log

Mao Nishiguchi[1,a)] Fujio Toriumi[1,b)] Masanori Takano[1,2,c)]

Received: November 6, 2019, Accepted: July 7, 2020

Abstract: Recently, it has become an important issue to reduce the risk of crimes, such as cyber-bullying and cyber-luring. One traditional approach is to build a model that automatically detects high-risk messages and users. These models mainly use a conversation corpus. However, there are cases where it is difficult to use a conversation corpus in the real world. Furthermore, it has been significantly increasing in recent years of the risk on the media that makes it possible to meet an unspecified number of people. The purpose of this study is to develop a model that estimates groups with high online-risk using only the action log. As a result of the construction experiments of the two-class classification model using actual data, we succeeded in building a relatively high performance model with Macro-F1 value of 0.883. In addition, we have obtained some interesting findings such as "Groups which users with specific network structures belong to have high crime risks."

Keywords: social networks, online-risk detection, cyber-bullying, cyber-luring

1. はじめに

Twitter や Facebook などのソーシャルメディアと呼ばれるサービスの利用率は年々増加傾向にある．総務省の調査によると，平成28年には調査対象者の71.2%がソーシャルメディアを利用している [31]．ソーシャルメディアが普及するにつれて，利用者，特に未成年者のネットリスクも高まっている．ソーシャルメディアを含むインターネット上のリスクとしては，誘い出し [13] や情報漏洩，ネット依存 [33]，ネット炎上 [32] など，様々なリスクが指摘されている [23]．本稿におけるネットリスクは，そ

[1] 東京大学
The University of Tokyo, Chiyoda, Tokyo 101–0062, Japan
[2] 株式会社サイバーエージェント
CyberAgent.inc, Chiyoda, Tokyo 101–0062, Japan
[a)] nishiguchi@crimson.q.t.u-tokyo.ac.jp
[b)] tori@sys.t.u-tokyo.ac.jp
[c)] takano_masanori@cyberagent.co.jp

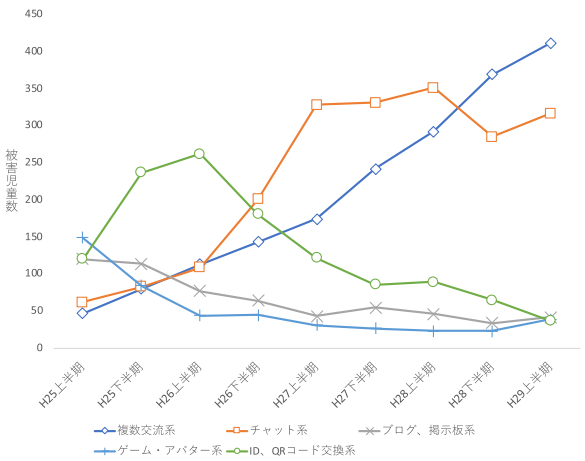

図 1 主要サイト種別の被害児童数の推移（文献 [27]，p.7 の図表を一部修正）

Fig. 1 Trends in the number of child victim by major site types. This figure is a partial modified version of page 7 of Ref. [27].

の中でもインターネットや電子メールのような電子通信を利用したいじめ（cyber-bullying）[20]，嫌がらせや脅迫行為（cyber-stalking）[8]，誘い出し（cyber-luring）[13] が起こる可能性と定義する．これらのネットリスクは，複数人での交流，すなわちグループチャットが可能なメディアにおいて特に増加している（**図 1**）[27]．複数交流系メディアのように，不特定多数の人々と気軽に交流できることは，様々なメリットをもたらす一方で，誰もが容易に加害者または被害者となる危険性もあわせ持つ [34]．ネットリスクを軽減するための仕組みを整えることは重要な課題である．

上記のような課題を解決するためのアプローチの 1 つとして，主にソーシャルメディア内に蓄積されたデータセットを利用して，ネットリスクを自動検知する取り組みがある [11]．過去多くの研究で，ユーザ間の会話コーパスを主な入力とした，メッセージやユーザのリスクの高さを推定する手法が提案されており，高性能なモデルの開発に成功した事例も報告されている [4], [5], [6], [7], [9], [16], [19], [26]．

既存研究で示されるように，会話コーパスは非常に強力な説明変数となりうるものの，通信の秘密や個人情報保護の観点から，データ解析への利用自体が困難なケースも存在する [14]．モデルの汎用性という観点からは，ユーザのデモグラフィック属性やメディア内のフォロー数のような，大半のメディアが利用可能なメタデータを用いることが望ましい．また，前述のように近年は複数交流形メディアのネットリスクが高まっており，グループチャットがきっかけとなる事犯も増加しているため，特定の会話や人物を検知するだけでは不十分であると考える．

本研究では，ネットリスクを軽減するための汎用的なシステムの開発の一環として，会話コーパスを用いずに，メタデータのみを利用した，ソーシャルメディア内に存在するネットリスクの高いグループを識別するモデルの構築を試みる．グループレベルでのネットリスクの識別を試みた研究は，我々が調べた限りまだない．そこで我々は，まずは汎用的な分類手法を用いたシンプルなモデルを構築することで，メタデータのみを利用した識別が可能であるか，また，どのような要因がモデルに寄与するかを明らかにする．

分析の手順としては，まず，ネットリスクが高いグループとそうでないグループを推定する 2 クラス分類問題を設定する．説明変数には，多くのソーシャルメディアで利用可能な情報である，参加ユーザ数やメッセージ投稿数のようなグループに付随する属性，および各ユーザの所属グループを表す関係ネットワークの構造情報を保存した分散表現ベクトルを利用する．分類アルゴリズムには，解釈性が高く高性能なモデルを構築可能な Random Forest [12] を採用する．次に，構築されたモデルを解釈することで，ネットリスクの高いグループにのみ特徴的に出現する要因を明らかにする．

本稿の主な貢献は以下のとおりである．

- 本研究は，会話コーパスを用いずに，グループレベルのネットリスク検知を試みた初めての研究である．
- シンプルなグループ属性情報，および関係ネットワーク構造情報のみを説明変数として，高性能なモデルを構築可能であることを示した．
- グループレベルでのリスク検知においては，特定のネットワーク構造情報が意味を持つ一方で，グループへの参加ユーザ数や投稿数のような，いわゆる注目度の高さを示す情報は直接的には寄与しない可能性があるなど，いくつかの興味深い知見を得た．

次章では，ネットリスク検知に関連する既存研究を紹介する．続く 3 章では，実験で使用するデータの概要や，設定した分類タスクについて述べる．4 章では，ネットリスク検知モデルの構築手法について説明し，5 章で実験結果について議論する．最後に，本研究の結論および今後について述べる．

2. 関連研究

2.1 ネットリスクの高いメッセージの検知

ネットリスク検知の領域で，最も活発に取り組まれてきたテーマは，あるメッセージのネットリスクの高さを推定するモデルを構築することである．Reynolds and Edwards [19] は，Q&A 型サイトのデータを使用して，あるメッセージが cyber-bullying 行為に該当するかどうかを推定する分類モデルを構築した．彼らは，事前に選択された非常に限定的な単語からなる Bag of Words（BoW）のみを入力として，C4.5 アルゴリズム [17] やインスタンスベース学習器 [1] などを適用することで，比較的高性能なモデルを構築することに成功した．ほかにも，会話コーパスに基づく特徴量を

主な入力として，cyber-bullying や cyber-stalking に該当するメッセージを検知するモデルの構築手法が提案されてきた [4], [5], [6], [7].

会話コーパスから得られた情報は，非常に強い説明力を持つものの，BoW のような単純な特徴量のみを入力とするモデルの精度には限界があることも指摘されている [11]. また，日本における主要ソーシャルメディアの 1 つが個人間のテキストメッセージの情報を利用しないと宣言しているように [14]，会話コーパスはプライバシ保護の観点から利用できないケースが存在する．汎用的なモデルの構築のためには，会話内容以外の情報を入力とすることが望ましい．

2.2 ネットリスクの高いユーザの検知

前節で紹介したような，メッセージに着目したモデルは，リスクの発生を迅速かつ正確に検知するという意味では大きな価値がある．しかしながら，これらのモデルはメッセージが投下された後に効果を発揮するものであるため，リスクの発生を未然に防ぐことは困難である．リスク軽減のためには，潜在的な加害者と被害者の出会い，すなわちリスクの高い出会い自体を防ぐことも考える必要がある．

高リスクな出会いを防ぐために，近年では高リスクなユーザを推定する取り組みもさかんである．Nahar ら [16] は，チャットログをユーザ–ユーザ関係ネットワークとして定義し，それを分析することで cyber-bullying の被害者と加害者を推定する方法を提案した．関係ネットワークがネットリスクの検知に寄与することは，他の研究でも示唆されている [2], [10], [22]．ほかにも，会話の時系列を考慮した検知モデル [15] や，経験的に定義された説明変数を利用した性犯罪者の検知 [3]，いじめの役割を推定する研究 [18] なども行われている．

高リスクユーザを推定することは，ネットリスクの軽減のために有益な手段である．しかしながら，前述のように，近年はグループでの交流が可能なメディアメディア上でのネットリスクが顕著に高まっており，事犯のきっかけがグループチャットであるケースも増加している．また，森田ら [30] は，あるいじめとは集団の問題であると定義しており，戸田ら [29] は，インターネット上のいじめは急激に集団化すると指摘している．したがって，ネットリスクの軽減のためには，ユーザレベルでの推定のみならず，グループレベルでの推定も必要であると考えている．

そこで我々は，ネットリスクを軽減するための新たなアプローチとして，ネットリスクの高いグループの推定に焦点を当てる．ネットリスクに関連する研究の中で，リスクの高いグループの推定を主目的とした研究は，我々が調べた限りまだない．

3. 使用データおよび分類タスク

3.1 使用データ

使用するデータは，株式会社 7gogo [*1]が運営する「755」という複数交流系ソーシャルメディアにおいて蓄積されたユーザの行動履歴データセットである．提供データは，ユーザ ID は仮名化されており，またすべてのコミュニケーションデータは公開されたスペースでのやり取りである．また，データは，株式会社 7gogo から研究を受託した親会社の株式会社サイバーエージェント[*2]が，執筆者の所属する東京大学の鳥海研究室と共同研究を行うことで提供された[*3]．なお，本稿では統計的分析のみを実施しているため，個人情報保護委員会の Q&A [28] の A2-5 に記載されているとおり，今回は利用目的の特定が不要な利用である[*4].

「755」は，ユーザが「トーク」と呼ばれるオンライン交流空間を自由に作成可能であり，トークに招待されたユーザはその空間内にテキストや画像などを投稿することが可能となる．「トーク」内に投稿されたコンテンツは，トークに招待されていないユーザでも自由に閲覧およびコメントを送信することができる．「755」ではまた，投稿されたコメントが他社サービスへの誘い出し行為や，いじめに繋がる誹謗中傷や脅迫行為に該当すると運営が判断（以下，この判断を NG と呼ぶ）した場合，当該コメントは削除される．

提供データの期間は 2015 年 1 月 4 日から 2015 年 3 月 31 日までの約 3 カ月間であり，「トーク」の作成や「トーク」へのメッセージ投稿，NG ラベル付きコメント投稿に加え，「トーク」に対してポジティブな感情を表現する「いいね」ボタンの押下ログやユーザ間のフォロー関係データなどが使用可能である．本研究では，3 人以上が参加している「トーク」をグループと定義する．

「755」と日本における主要なソーシャルメディア（LINE や Twitter など）との間の違いは，大きく 2 つある．まず，グループ内会話がオープンであることである．「755」上で 3 人以上の複数人が参加する「トーク」は，実質的にグループチャットのサービスを提供することになるが，主要なソーシャルメディア上のグループチャットでは，基本的にグループに招待されたユーザのみ閲覧可能であるという制限機能が付加されている．次に，利用ユーザ数に違いがある．主要なソーシャルメディアは日々数千万人が利用しているサービスであるのに対し，「755」は提供データ期間中の利用ユーザ数は 1 日あたり数十万人である．したがって，今回の実験で得られた結果が，他のソーシャルメディ

[*1] http://www.7gogo.co.jp/
[*2] https://www.cyberagent.co.jp/
[*3] 研究開発部門が親会社である株式会社サイバーエージェントにのみ存在しているため，親会社を経由している．
[*4] すなわち，755 のプライバシポリシ [24]・利用規約 [25] への記載の有無を問わない．

表 1 各クラスのグループ数および割合
Table 1 Number and ratio of groups in each class.

クラス	グループ数	グループ数割合
RISK	4,561	0.11
SAFE	37,519	0.89

アデータにおいてもあてはまるかどうかについては議論できない．メディア間の違いに関しての検証は今後の課題とする．

3.2 分類タスク

本稿ではネットリスクの高いグループを分類するモデルを構築する．設定した分類タスクは，データ提供期間の初日である 1 月 4 日以降に作成されたグループのうち，3 月 1 日から 3 月 31 日の間に NG コメントが発生したグループ集合と，提供データ期間中 1 度も NG コメントが発生しなかったグループ集合を識別する 2 クラス分類問題である．1 月 4 日から 2 月 28 日の間に NG コメントが発生したグループは対象から除外している．また，特徴分析のためにはある程度のデータ量が蓄積されていることが望ましいため，グループ内への投稿やコメントの総数が 10 件以上であったグループのみを対象とする．以下では，NG コメントが発生したクラスを RISK クラス，他方のクラスを SAFE クラスと呼ぶ．各クラスのグループ数およびグループ数割合を**表 1** に記載する．クラス比は 11 対 89 と，比較的不均衡なデータである．

未然検知という観点では，予測を目的としたタスク，たとえば 2 月 28 日までの情報のみを用いて 3 月 1 日以降を予測するなど，を設定するのが妥当である．しかしながら，戸田ら [29] が指摘するように，ネットいじめが急激に集団化するのであれば，直前の行動がモデルに寄与するのではないかと考える．実際，我々が行った事前分析においても，グループの作成日に NG コメントが頻出するなどの傾向が確認された．そこで今回はリアルタイムの検知，いわゆるオンライン予測システムの開発を想定して，RISK クラスにおいては，NG コメントが発生する直前までのデータを説明変数の作成期間として利用する．

一方の SAFE クラスにおいては，RISK クラスのグループが取り得る最長の期間である 3 月 31 日までのデータを利用して説明変数を作成する．この設定では，使用するデータ期間に違いが生じてしまうが，時間的情報および NG コメントの出現有無に関する情報を含む説明変数は作成しないことから，データリークは発生しないため，大きな問題にはならないと考える．説明変数の詳細については次章で述べる．

4. 高リスクグループ推定モデル

提案モデルには，2 つの異なる観点から作成された，多くの複数交流系ソーシャルメディアで利用可能な説明変数を入力として与える．1 つは投稿数や参加ユーザ数のようなグループ属性，もう 1 つはユーザとグループの所属関係により定義されるユーザ–グループ関係ネットワークの構造情報から抽出された分散表現である．

表 2 各グループ属性変数の基本統計量
Table 2 Basic statistics for each group attribute variable.

説明変数	最小値	最大値	平均値	標準偏差
参加ユーザ数	3	1,531	16.8	40.5
投稿数	10	144,117	601.0	2,611.8
いいね数	0	7,977	0.5	40.9
過去 NG ユーザ数	0	142	0.2	2.1
最大フォロー数	0	273	0.6	4.8
最大フォロワー数	0	276	0.5	4.3

4.1 グループ属性

グループ属性としては以下の 6 つの変数を準備する．

(1) 参加ユーザ数：グループに招待されたユーザおよびコメント投稿を行ったユーザの総数．
(2) 投稿数：グループ内に投稿されたメッセージおよびコメント投稿の総数．
(3) いいね数：グループに対する「いいね」ボタンの押下数．
(4) 過去 NG ユーザ数：1 月 4 日から 2 月 28 日の間に，755 内に存在するいずれかの「トーク」において，1 件以上の NG コメントを投稿したユーザの参加人数．
(5) 最大フォロー数：グループ内にメッセージまたはコメントを投稿したユーザのなかで，最も多くのフォローを行っているユーザのフォロー数．
(6) 最大フォロワー数：グループ内にメッセージまたはコメントを投稿したユーザのなかで，最も多くフォローされているユーザのフォロワ数．

なお，最大フォロー数とフォロワ数は，グループ内の会話がどれだけ多くのユーザにリーチするかを示す近似値として採用している．各変数の基本統計量を**表 2** に記載する．

4.2 ネットワーク構造

Nahar ら [16] の研究成果から，関係ネットワークの構造情報は本研究においても有効ではないかとの仮説のもとで，ユーザと所属するグループという 2 種類のノードからなる 2 部の有向ネットワークを関係性ネットワークとして用いる．ネットワークの基本統計量を**表 3** に示す．平均出次数はユーザ 1 人あたりの所属グループ数，平均入次数は 1 グループあたり参加人数と同義である．表から，比較的大規模かつ疎なネットワークであることが分かる．また，今回はユーザ数が 30 万人以上であるため，ネットワークの構造情報をそのまま説明変数として利用するのは計算コスト

の観点から困難である．近年の主要なソーシャルメディアでは，ネットワークはさらに大規模なものとなるため，今後の拡張性を考えてもネットワーク構造を何らかの低次元ベクトル（分散表現）に埋め込む手法が必要となる．本研究では，有向2部ネットワークに適した代表的な分散表現獲得手法の1つである，Large-scale Information Network Embedding（LINE）(2nd) [21] を適用する．

LINE(2nd) の学習プロセスを説明する前に，本章で扱うネットワークを定義する．ノード集合 V，および有向エッジ集合 E が与えられたとき，有向ネットワークは $G = (V, E)$ と定義される．各エッジ $e \in E$ はノードの順序付きペア $e = (u, v)$ であり，ノード間の接続の強さを表す重み $w_{uv} > 0$ を持つ．LINE(2nd) の目的は，各ノード $v \in V$ を，ある低次元空間 R^n で表現することである．ただし，$n \ll |V|$ である．

LINE(2nd) では，あるノードは他のノードの文脈という役割が与えられる．そして，文脈にわたり類似の分布を有するノードは類似していると仮定する．LINE(2nd) は，この仮説を経験的に表現した確率分布 $\hat{p}(\cdot|v_i)$ と，分散表現ベクトルの内積により得られる確率分布 $p(\cdot|v_i)$ との間の差を最小化するように学習する．LINE(2nd) が解く目的関数は式 (1) で定式化される．

$$O = \sum_{i \in V} \lambda_i d\left(\hat{p}(\cdot|v_i), p(\cdot|v_i)\right) \tag{1}$$

ここで，$d(\cdot, \cdot)$ は2つの確率分布間の距離であり，相対エントロピーによって算出される．λ_i はネットワーク内のあるノード i の重要度を表しており，$\lambda_i = d_i$，$d_i = \sum_{k \in N(i)} w_{ik}$ である．ここで $N(i)$ はノード v_i の近傍である．

経験的確率分布は式 (2) で，分散表現の内積から得られる確率分布は式 (3) で定義される．

$$\hat{p}\left(v_j|v_i\right) = \frac{w_{ij}}{d_i} \tag{2}$$

$$p\left(v_j|v_i\right) = \frac{\exp\left(\vec{u}_j'^T \cdot \vec{u}_i\right)}{\sum_{k=1}^{|V|} \exp\left(\vec{u}_k'^T \cdot \vec{u}_i\right)} \tag{3}$$

ここで，$\vec{u}$ はノードの役割の時の v_i の表現であり，$\vec{u}_i'$ は特定の文脈として扱われる時の v_i の分散表現を意味する．$|V|$ は文脈の数である．

LINE(2nd) はまた，negative sampling およびエッジサンプリングにより，高速かつ正確な学習を実現している．アルゴリズムの詳細は文献 [21] を参照されたい．

本研究においては，埋め込み後の次元数 n は事前パラメータとし，分類モデルの性能が最も高くなった数を採用する．詳細は次節で述べる．なお，モデルの説明変数としては，グループに相当するノードの側の分散表現ベクトルを利用する．

4.3 実験設定および評価方法

分類手法には，高性能かつ可読性の高い Random Forest [12] を利用する．ノードの分割基準には gini 係数を採用し，分割時に探索する説明変数はランダムサンプリングにより $\sqrt{m}$ 個の変数を都度抽出する．ここで，m は説明変数の総数である．

モデルの評価には10分割交差検証法を採用し，Macro-F1値および各クラスの F1 値，Precision，Recall により評価する．すべての評価値は 10 シード分の結果の平均値を記載する．また，表 1 で示したように，今回はクラス比率が不均衡であり，そのまま手法を適用しても高性能なモデルを構築することは難しい．したがって，訓練データは RISK クラスのグループ数をランダムにオーバサンプリングして，クラス比が1対1となるように事前に調整する．

事前パラメータに関しては，LINE(2nd) の埋め込み後の次元数は 10 から 100 を 10 刻みで実行し，交差検証法により最も評価値が高い次元数とする．Random Forest を構成する学習器の数，木の深さ上限，そして各葉ノードが分岐を行うための最小インスタンス数はそれぞれ，1 から 5，1 から 10，5 から 50 を探索範囲とし，これらのパラメータも交差検証法による評価値に基づき決定する．

また，グループ属性情報とネットワーク構造情報それぞれの，単独でのモデル性能を確認するため，一方の説明変数のみを入力としたモデルの評価も行う．以下で計算実験の結果について議論する．

5. 実験結果

以上の設定により，計算実験を実施した．採用された事前パラメータの一覧を**表 4** に記載する．

5.1 モデル評価値

得られたモデルの評価値を**表 5** に示す．表中のグループ属性とは，グループ属性変数のみを入力としたモデルを意

表 3 ネットワークの基本統計量
Table 3 Basic statistics of network.

	件数
グループ数	42,080
ユーザ数	349,074
エッジ数	706,203
平均出次数	2
平均入次数	17

表 4 事前パラメータの値
Table 4 Value of hyper-parameters.

	事前パラメータ	値
LINE(2nd)	次元数	90
Random Forest	学習器の数	4
	深さ上限	2
	最小インスタンス数	5

表 5 テストデータに対する評価値
Table 5 List of evaluation values for test data.

	指標	グループ属性	分散表現	全変数
	Macro-F1	0.864	0.518	**0.883**
RISK	F1	0.751	0.161	**0.787**
	Precision	**1.000**	0.149	0.999
	Recall	0.602	0.196	**0.649**
SAFE	F1	0.977	0.875	**0.979**
	Precision	0.954	0.897	**0.959**
	Recall	**1.000**	0.856	**1.000**

味し，分散表現とは分散表現のみを入力としたモデル，全変数とはグループ属性変数と分散表現の両方を入力としたモデルを意味する．各指標のベスト結果は太字にしている．

結果を評価すると，まず，すべての評価値において全変数モデルがベストまたは次点の結果であった．全変数モデルの Macro-F1 値は 0.883 と高性能であり，クラス比率が低く分類が難しい RISK クラス側の F1 値も 0.787 と高い値となっている．この結果から，少なくとも本研究で使用したデータに対しては，十分に意味のあるモデルであるといえる．また，グループ属性のみを用いたモデルでも同様に高性能なモデルが構築可能であった．一方，分散表現のみを入力としたモデルは，特に RISK クラス側の分類能力が低く，単体では十分な性能であるとはいえない．全変数モデルの性能はグループ属性モデルを上回っていることから，分散表現はグループ属性と組み合わせることで精度向上に多少寄与しているものの，Macro-F1 値でわずか 0.02 ポイントの向上に留まっている．これは，表 3 で示したように，ネットワークが疎であることに原因があると考える．LINE(2nd) は，直観的には同じユーザが所属しているグループほど分散表現空間上の距離が近くなりやすいという性質を持つが，ユーザの所属グループ数が 1 ユーザあたり約 2 つと少ないために，意図した表現空間を取得できなかった可能性が考えられる．今後，疎なユーザ–グループ関係ネットワークに対しては，媒介中心性のような他の指標に代替するか，ネットワークの密度を高めるなどの工夫を行うことで，適切な説明変数を模索していく．

5.2 RISK クラスの特徴分析

続いて，モデルの解釈を行っていく．各変数のモデル内重要度を算出し，降順に並べ替えてグラフ化したものが**図 2** である．重要度は，学習器の構築の際の分割により減少した gini 係数値と，分割後のノードに該当するグループ数の積によって算出された値の平均値である．図中に存在しない変数は，モデルに 1 度も採用されていない変数，すなわち重要度が 0 であったことを意味する．図中の 32 次元および 23 次元は，ともに分散表現の次元であり，番号に意味はない．図から，グループ属性変数が大きく寄与していることが分かる．特に，グループに参加するユーザのフォローおよびフォロワ数の最大値や，いいねボタンの押下数が，非常に重要な意味を持つことが明らかとなった．一方，過去 NG ユーザ数は，クラス定義と最も関連すると考えられる変数ではあるが，モデルにはまったく寄与していないことが分かる．参加ユーザ数や投稿数も，確率論的には，多ければ多いほど NG コメントの出現確率も上昇するが，当該データにおいては大きな意味を持たない変数であったことは興味深い．分散表現の重要度は僅かではあるものの，グループ属性のみモデルより高性能なモデルを構築するために寄与していることが確認できる．

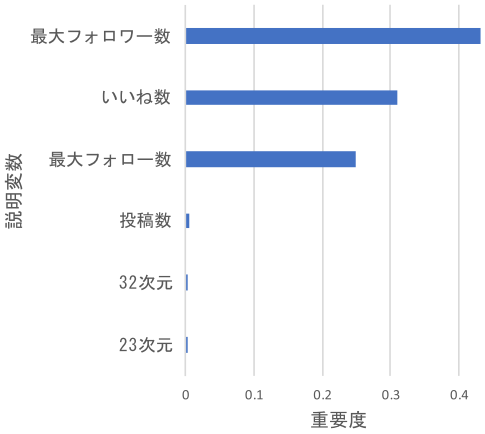

図 2 変数重要度
Fig. 2 Importance of features.

表 6 RISK クラスに特徴的なルール
Table 6 Characteristic rules for RISK class.

ルール	グループ数	gini 係数
{最大フォロワー数 ≥ 1}	11,454	0.0
{いいね数 ≥ 1}	4,122	0.0
{23 次元 ≥ −0.19, 最大フォロー数 ≥1}	11,262	0.0
{23 次元 ≥ −0.19, 投稿数 ≥ 62}	248	0.207
{いいね数 <1, 最大フォロワ数 ≥1}	8,543	0.0

最後に，RISK クラスに特徴的であったルールを考察する．**表 6** は，各学習器を構成する分岐ルールのうち，RISK クラスに該当するルールのみを抽出したものである．表中のグループ数とは，ルールに該当するグループ数を意味する．ここで，この表のグループ数はオーバサンプリング後の数であり，総数は SAFE クラスと同じ 37,519 であることに注意されたい．また，ルール内のカンマは AND 条件を意味する．すなわち，{いいね数 < 1, 最大フォロワー数 ≥ 1} とは，いいね数が 1 未満，かつ，最大フォロワー数が 1 以上というルールを表す．表から，変数重要度の上位 4 つの変数は，値が大きいほど RISK クラスに該当するルールであることが分かる．閾値はサービスの規模に依存するが，いいね数が多いことや，フォロワー数が相対的に多いユーザの参加は，多くのユーザの注目を浴びているグループといい換えることができる．このようなグループは，多様なユーザで構成されている可能性が高いため，ユーザ同

士の衝突や公序良俗に反する発言が発生しやすいのではないかと考えられる．また，分散表現空間の特定の領域に位置付けられるユーザの存在，すなわち，特定のネットワーク構造を持つユーザがネットリスクを高めている可能性がある．こうしたユーザは，意識的か否かにかかわらず，ネットリスクを高めるきっかけとなってしまっている可能性がある．このようなユーザが所属するグループへの参加には注意が必要である．

6. おわりに

本研究では，ソーシャルメディア上のネットリスクの高いグループを推定する分類モデルを構築した．実データを使用した計算実験を行うことで，会話コーパスを用いずとも，シンプルなグループ属性情報，および関係ネットワーク構造情報のみを説明変数として高性能なモデルが構築可能であることを確認した．また，モデルを読み解くことで，多くの注目を集めているユーザが所属するグループや，特定のユーザ–グループ関係ネットワーク構造を持つユーザが所属するグループのネットリスクが高まる可能性があることなど，いくつかの有用な知見を得た．なお，これらの知見は，あくまで本研究で使用したデータにのみ当てはまるものであることに注意されたい．

今後は，他のソーシャルメディアおける類似の分析を実施することで，より汎用的な知見の獲得を目指すとともに，グループレベルでのネットリスクの推定という技術の実社会への具体的な応用方法を模索し，より安全なサイバー空間を作るための努力を行っていく．

謝辞 本研究は RISTEX「未成年者のネットリスクを軽減する社会システムの構築」プロジェクトの助成を受けた研究である．また，貴重なデータをご提供いただいた株式会社 7gogo および株式会社サイバーエージェントの皆様に感謝申し上げます．

参考文献

[1] Aha, D.W., Kibler, D. and Albert, M.K.: Instance-based learning algorithms, *Machine Learning*, Vol.6, No.1, pp.37–66 (1991).
[2] Asatani, K., Kawahata, Y., Toriumi, F. and Sakata, I.: Communication based on unilateral preference on Twitter, *Proc. 10th International Conference on Social Informatics* (*SocInfo 2018*), pp.54–66 (2018).
[3] Cardei, C. and Rebedea, T.: Detecting sexual predators in chats using behavioral features and imbalanced learning, *Natural Language Engineering*, Vol.23, No.4, pp.589–616 (online), DOI: 10.1017/S1351324916000395 (2017).
[4] Dadvar, M., Jong, D.F., Ordelman, R. and Trieschnigg, D.: Improved cyberbullying detection using gender information, *Proc. 12th Dutch-Belgian Information Retrieval Workshop* (*DIR 2012*), University of Ghent (2012).
[5] Dinakar, K., Jones, B., Havasi, C., Lieberman, H. and Picard, R.: Common sense reasoning for detection, prevention, and mitigation of cyberbullying, *ACM Trans. Interactive Intelligent Systems* (*TiiS*), Vol.2, No.3, p.18 (2012).
[6] Dinakar, K., Reichart, R. and Lieberman, H.: Modeling the detection of Textual Cyberbullying, *The Social Mobile Web*, Vol.11, No.2, pp.11–17 (2011).
[7] Frommholz, I., Al-Khateeb, H.M., Potthast, M., Ghasem, Z., Shukla, M. and Short, E.: On textual analysis and machine learning for cyberstalking detection, *Datenbank-Spektrum*, Vol.16, No.2, pp.127–135 (2016).
[8] Goodno, N.H.: Cyberstalking, a new crime: Evaluating the effectiveness of current state and federal laws, *Mo. L. Rev.*, Vol.72, p.125 (2007).
[9] Hirano, Y., Toriumi, F., Tashiro, M. and Eguchi, K.: Finding Minors Faced with Online Risk, *Journal of Transformation of Human Behavior under the Influence of Infosocionomics Society*, Vol.3 (2018).
[10] Huang, Q., Singh, V.K. and Atrey, P.K.: Cyber Bullying Detection Using Social and Textual Analysis, *Proc. 3rd International Workshop on Socially-Aware Multimedia*, pp.3–6, ACM (2014).
[11] Huang, Q., Singh, V.K. and Atrey, P.K.: On cyberbullying incidents and underlying online social relationships, *Journal of Computational Social Science*, Vol.1, No.2, pp.241–260 (2018).
[12] Liaw, A., Wiener, M., et al.: Classification and regression by randomForest, *R news*, Vol.2, No.3, pp.18–22 (2002).
[13] Lievens, E.: New criminal provisions for online solicitation for sexual purposes and cyberluring, *IRIS* (*ENGLISH ED. ONLINE*), No.7, p.6 (2014).
[14] LINE 株式会社：「サービス向上のための情報利用に関するお願い」について，よくあるご質問および詳細情報（オンライン），入手先 ⟨https://terms.line.me/line_communication_privacy⟩（参照 2020-06-01）.
[15] Liu, D., Suen, C.Y. and Ormandjieva, O.: A Novel Way of Identifying Cyber Predators, arXiv preprint arXiv:1712.03903 (2017).
[16] Nahar, V., Li, X. and Pang, C.: An effective approach for cyberbullying detection, *Communications in Information Science and Management Engineering*, Vol.3, No.5, p.238 (2013).
[17] Quinlan, J.R.: *C4.5: programs for machine learning*, Elsevier (2014).
[18] Raisi, E. and Huang, B.: Cyberbullying detection with weakly supervised machine learning, *Proc. 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp.409–416, ACM (2017).
[19] Reynolds, K., Kontostathis, A. and Edwards, L.: Using machine learning to detect cyberbullying, *2011 10th International Conference on Machine Learning and Applications and Workshops* (*ICMLA*), Vol.2, pp.241–244, IEEE (2011).
[20] Smith, P.K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S. and Tippett, N.: Cyberbullying: Its nature and impact in secondary school pupils, *Journal of Child Psychology and Psychiatry*, Vol.49, No.4, pp.376–385 (2008).
[21] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J. and Mei, Q.: LINE: Large-scale Information Network Embedding, *WWW*, ACM (2015).
[22] Toriumi, F., Nakanishi, T., Tashiro, M. and Eguchi, K.: Analysis of User Behavior on Private Chat System, *Jour-*

nal of Transformation of Human Behavior under the Influence of Infosocionomics Society, Vol.3, pp.23–28 (2016).
[23] 伊藤賢一：スマートフォン時代における青少年のリスク構造–群馬県前橋市調査より (2016).
[24] 株式会社 755：プライバシーポリシー（オンライン），入手先 ⟨https://755.themedia.jp/pages/245366/privacy⟩（参照 2020-06-01）.
[25] 株式会社 755：利用規約（オンライン），入手先 ⟨https://755.themedia.jp/pages/245361/rules⟩（参照 2020-06-01）.
[26] 吉田俊和，大西彩子，黒川雅幸，三島浩路，本庄 勝，吉田恵子：J004 携帯電話によるネットを媒介とした「いじめ」防止システムの開発：中学生・高校生を対象とした実用的なシステムづくり（自主シンポジウム），日本教育心理学会総会発表論文集第 52 回総会発表論文集，一般社団法人日本教育心理学会，pp.100–101 (2010).
[27] 警察庁：平成 29 年上半期におけるコミュニティサイト等に起因する事犯の現状と対策（オンライン），入手先 ⟨https://www.npa.go.jp/cyber/statics/h29/H29_siryou.pdf⟩（参照 2020-06-01）.
[28] 個人情報保護委員会：「個人情報の保護に関する法律についてのガイドライン」及び「個人データの漏えい等の事案が発生した場合等の対応について」に関する Q&A（オンライン），入手先 ⟨https://www.ppc.go.jp/files/pdf/180720_APPI_QA.pdf⟩（参照 2020-06-01）.
[29] 戸田有一，青山郁子，金綱知征：ネットいじめ研究と対策の国際的動向と展望，教育と社会研究，Vol.23, pp.29–39 (2013).
[30] 森田洋司，清永賢二：いじめ：教室の病い，金子書房 (1986).
[31] 総務省：平成 28 年情報通信メディアの利用時間と情報行動に関する調査報告書（オンライン），入手先 ⟨http://www.soumu.go.jp/main_content/000564530.pdf⟩（参照 2020-06-01）.
[32] 田中辰雄，山口真一：ネット炎上の研究，勁草書房 (2016).
[33] 樋口 進：ネット依存症，Vol.894, PHP 研究所 (2013).
[34] 文部科学省：「ネット上のいじめ」に関する対応マニュアル・事例集（学校・教員向け）（オンライン），入手先 ⟨http://www.mext.go.jp/b_menu/houdou/20/11/08111701/001.pdf⟩（参照 2020-06-01）.

西口 真央

2013 年大阪府立大学大学院経済学研究科博士後期課程修了．博士（経済学）．複数の IT ベンチャー企業を経て，2018 年から東京大学大学院工学系研究科特任研究員．専門はデータマイニングや機械学習技術のビジネス応用．人工知能学会，オペレーションズ・リサーチ学会，経営情報学会，日本マーケティング・サイエンス学会各会員．

鳥海 不二夫 （正会員）

2004 年東京工業大学大学院理工学研究科機械制御システム工学専攻博士課程修了．博士（工学）．同年名古屋大学情報科学研究科助手，2007 年同助教，2012 年東京大学大学院工学系研究科准教授．計算社会科学，人工知能技術の社会応用等の研究に従事．情報法制研究所理事．人工知能学会，電子情報通信学会，日本社会情報学会，AAAI 各会員．「科学技術への顕著な貢献 2018（ナイスステップな研究者）」に選定．

高野 雅典

2009 年名古屋大学大学院情報科学研究科博士課程修了．博士（情報科学）．システムインテグレータを経て，株式会社サイバーエージェントに勤務．スマートフォンゲームの開発・運用に携わった後，現在はメディアサービスのデータ分析と計算社会科学研究に従事．専門は計算社会科学・複雑系科学．人工知能学会，行動計量学会各会員．