Vol.2011-CSEC-55 No.2
2011/12/5

ソーシャルメディアと履歴書情報の照合を通じた個人の言動の推定

奥野智孝[†]　市野将嗣[†]
久保山哲二[††]　吉浦裕[†]

近年，多様な個人情報がネットワーク上に流通している．同一人物に関する複数の情報を入手することで，単独の情報からでは分からなかった情報が明らかになり，予期せぬプライバシー侵害につながる懸念がある．本研究ではこの危険性を明らかにするために，問題の代表例としてバックグラウンドチェックと呼ばれる雇用前の身辺調査を例に挙げ，ソーシャルメディアのプロフィールが匿名化されていても，履歴書の情報を基にソーシャルメディアのコンテンツの特徴を分析することで，採用希望者のアカウントを特定できることを示した．これにより，履歴書の情報とソーシャルメディアで開示された情報を統合し，個人の言動を調査することができる．

Content-based De-anonymization of Tweets

Tomotaka Okuno[†]　Masatsugu Ichino[†]
Tetsuji Kuboyama[††]　and　Hiroshi Yoshiura[†]

Various types of personal information about individuals are accessible through the Web medias. Linking of the personal information obtained through multiple medias can lead to a serious violation of privacy. To address this problem, we developed a method to identify the author of the short messages of Twitter, known as *tweets*, by using the information from other medias.

1. はじめに

近年，ソーシャルメディアを通じた情報公開，不正アクセスによる企業データベースからの顧客情報の漏えい，ユビキタスネットワークにおける自動的な位置情報の収集など，多様な要因により様々な個人情報が流通している．これらの流通している個人情報を第三者が入手した場合，プライバシーの侵害につながる危険性がある．

このような問題を避けるために，それぞれの分野において様々な対策が検討，実施されている．例えばソーシャルメディアを利用する際に，コンテンツの公開範囲の設定やプロフィール情報の匿名化（名前などの情報を削除）といった対策を行う利用者は多くいる．米国の大学生に対して行われた調査では，54%の学生が公開範囲を設定していると報告している[1]．また，日本の大手 SNS（Social Networking Service）である「mixi」では，20 代前半のユーザのうち 60%以上が実名以外の名前で登録している[2].

一方，これらの対策にも関わらず，同じ人に関する複数の情報を入手することで，単独の情報からでは分からなかった情報が明らかになる可能性があり，個人の特定や個人の言動の把握など予期せぬプライバシー侵害につながる危険性がある．マイクロブログ「Twitter」にて問題発言をした人が SNS などの他の情報から特定され，炎上騒ぎになるという事例[3][4]は，複数の情報から個人が特定された事例の氷山の一角であると推測される．

本研究では，このような複数の情報の組み合わせから生じるプライバシー情報の漏えいの危険性を明らかにするために，問題の代表例として Background Check とよばれる採用希望者を対象とした雇用前の身辺調査を例に挙げる．この身辺調査の方法の 1 つに，採用希望者の提出した履歴書の情報を手がかりに，当該採用希望者がソーシャルメディアを通じて発信したコンテンツを特定する，というものがある．本研究では，ソーシャルメディアのプロフィールが匿名化されていても，コンテンツに含まれる言葉と履歴書との間接的な関わりを定量化することで，コンテンツの作者の特定が可能であることを示す．

2 章では先行研究について述べる．3 章では，本研究で取り上げる Background Check について述べ，4 章で具体的な課題について述べる．5 章ではその課題に対して従来手法を用いた方式について，6 章では改良した提案方式について述べる．7 章では結論と今後の課題について述べる．

[†] 電気通信大学
University of Electro-Communications
[††] 学習院大学
Gakushuin University

2. 先行研究

2.1 個々の分野におけるプライバシー情報保護

データベース，位置情報サービス，ソーシャルメディアの各分野でのプライバシー情報漏えいの危険性と，その問題への対策について述べる．

データベースの分野では，従来からアクセス制御や推論制御などが検討されている[5]．また，近年データベースを暗号化する手法[6]や秘密分散法を用いた手法[7]も検討されている．データベースの 2 次利用の際の保護対策として，テーブル内に QI（Quasi-Identifier:郵便番号や年齢など組み合わせることで個人を識別できる属性の集合）が同一であるレコードが k 個以上存在するように匿名化する k 匿名性[8]，またその問題を解決する l 多様性[9]，t 近傍性[10]が提案されている．

また位置情報の分野では，入手した情報を用いて攻撃者がユーザの目的地や自宅・勤務先を予測できることが指摘されている[11]．このような位置情報の累積によるプライバシー情報の漏えいを防ぐために，地理空間上への k 匿名性の適用[12]や，mix zones と呼ばれる事前に定義されたエリア内でサービスに公開する情報を変更する手法[13]などが提案されている．

ソーシャルメディアの分野では，大手 SNS である「Facebook」のユーザに対して調査した情報の公開範囲設定と情報漏えいについての実情[14]，またマイクロブログ Twitter の機能である Retweet（他のユーザの発言を引用形式により発信すること）による情報漏えいの問題を指摘したもの[15]などが報告されている．この問題に対して，公開範囲を記述したタグの添付によりそれぞれの日記の公開範囲を自動的に設定する手法[16]や，公開した文章からのプライバシー情報の漏えいに着目し，それらを検知して適切な表現に変換する自然言語情報の開示制御技術(DCNL)[17]などが検討されている．

2.2 情報の組み合わせを考慮したプライバシー保護

データベースの分野では，ユーザのプロフィールが匿名化された映画評価データベースに対して，大手映画情報サイトから入手できる情報を利用することでその匿名性を破ることが可能であるとの報告がある[18]．また，ソーシャルメディアのユーザ同士のつながりに着目し，ネットワークトポロジーを用いることで，異なる 2 種類のソーシャルメディアを利用する同一ユーザを特定する手法[19]やユーザ名を推定する手法[20]などが提案されている．他にも同一人物により Web 上に投稿されたテキストは，文章レベルで類似度を算出し照合することが可能であるとの報告がある[21]．

このように，複数のメディアから個人情報を入手することにより，個人の照合が可能であることが指摘されている．

3. Background Check について

Background Check とは，企業などが採用希望者に対して雇用前に行う職歴，学歴，犯罪歴などの身辺調査である．特に欧米では，被雇用者が何らかの事故や事件により第三者に被害を与えた場合，雇用側の雇用責任が問われ訴訟になることがあるため，この調査が行われることが多い．しかし，調査の仕方や目的によってはプライバシー侵害の問題となる可能性がある[22]．

近年，Background Check の調査対象が Web 上に拡大している．たとえば，採用希望者の氏名を Google 等の検索エンジンで検索することで，大学研究室のホームページや，（良くも悪くも掲載されるような事例があった場合は）ニュースサイト，また実名を公開している場合は個人のホームページを調べることができる．特に，SNS はプライベートな出来事や画像を公開し，それらの情報を友人と共有するユーザが多くいるため，採用希望者の人物像を簡単に知ることのできる情報源として注目され始めている．2008 年に米国で行われた調査によると，調査対象企業のうち 40.8%が採用希望者の SNS の個人ページを確認しているとの結果が得られた[23]．また，39%の企業が社員の SNS の個人ページを確認しているという報告もある[24]．

このようなソーシャルメディアを対象とした Background Check に対して採用希望者が取ることのできる対策は以下の 2 通りであると考えられる．

(1) コンテンツの公開制限を「友人まで公開」「プロテクト」など，限られたユーザのみ閲覧できるように設定
(2) プロフィールに登録している実名やメールアドレス，写真といった個人情報の非匿名化（削除など）

上記対策のうち，(1)は採用担当者が採用希望者の発信したコンテンツを見ることができなくなるため，安全ではある．しかし，ソーシャルメディアの醍醐味である多くの人とのコミュニケーションを阻害する．また，応募者の「友人」や「友人の友人」になり済ましてコンテンツを閲覧しようとする動きには対応できない．物理社会でなり済ましによる身辺調査があったこと[22]を踏まえると，同様の調査が Web 上で行われている可能性がある．

一方，(2)は，例えば友人が見れば分かる程度の情報に個人情報を置き換えることで，氏名検索のような単純な個人の特定は回避できる．また，顔画像の照合のような攻撃も回避できる．しかし，プロフィールから個人情報を削除することで匿名性が保たれていたとしても，ソーシャルメディアにはプロフィール以外の文章や写真などのコンテンツも多く蓄積されている．そのため，コンテンツの特性に基づいて採用者のアカウントが特定される可能性がある．特に，採用担当者は履歴書の情報や面接の内容といった予備知識を持つため，採用希望者のアカウントを推測できる可能性は高まる．

4. Web-based Background Check 技術の提案

4.1 WBC の目的

複数の情報を組み合わせたときに，さらなるプライバシー情報の漏えいが起きる危険性を明らかにし，社会に警鐘を鳴らすことを目的とする．そのために，採用希望者がソーシャルメディアのプロフィール情報を匿名化したとしても，発信された文章の特性を分析することで採用希望者のアカウントを特定し，Background Check を行うことが可能であることを示す．この目的のために，履歴書内の語句が文章内で直接使われていなくても，その語句と文章との関連性を定量化することで，採用希望者のアカウントを特定できることを示す．今後，このような Web（特にソーシャルメディア）を利用した Background Check システムのことを WBC(Web-based Background Check)と呼ぶ．

4.2 機能構成

WBC では，本人のものを含むソーシャルメディアのアカウントの候補が複数与えられた時に履歴書の内容に基づいて採用希望者のアカウントを特定する．これにより，ソーシャルメディアと履歴書の情報を統合して，採用希望者の日頃の言動を知ることが可能になる．

4.3 例題

採用希望者の履歴書と本人のものを含む Twitter のアカウントが複数与えられたとき，Twitter のつぶやき(tweet)と呼ばれる書き込みと履歴書の情報を照合して，採用希望者のアカウントを特定する．

(1) 履歴書

Adam，Bob（仮名）2 名の履歴書を用意した．履歴書に含まれる情報は以下の通りである．

[1] 現住所
[2] 学歴（所属研究室や研究内容，発表実績なども含む）
[3] 職歴
[4] 資格

(2) つぶやき

Adam，Bob のものを含む 5 名のアカウントのつぶやきを各 400 件用意した．それぞれのユーザの属性を表 1 に示す．表 1 から Adam，Bob が類似した属性を多く持つこと，同様に Bob と Charlie，Adam と Charlie もそれぞれ似た属性を持つことが分かる．また，Dave と Ellen は他の人と似た属性を持たない．

表 1 ユーザの属性

ユーザ名 (仮名)	出身大学	職場	専門分野
Adam	国立理工系 A 大学	通信系 C 社	ネットワークセキュリティ
Bob	私立総合 B 大学	通信系 C 社 → A 大学（准教授）	ネットワーク，マルチメディア
Charlie	国立理工系 A 大学	A 大学の学生	メディアセキュリティ
Dave	（不明）	D ゲーム会社	ゲームクリエイター
Ellen	（不明）	E プロダクション	歌手

5. WBC Basic

5.1 方式

つぶやきと履歴書をそれぞれ文書とみなし，それらの類似度を計算する．そして，類似度が最も高いつぶやきの著者を採用希望者と判断する．類似度の算出には，情報検索に用いられる代表的な検索モデルの一つであるベクトル空間モデル[25]を用いる．これは文書集合を多次元ベクトルによって表現し，ベクトル間の類似度を算出することにより文書間の類似検索を行うものである．ベクトルの表現には TFIDF[26]を，類似度の計算には余弦を用いた．なお，つぶやきは 1 件最大 140 文字と短いため，以下のように時系列に沿って 100 件ごとに区切り，それぞれを 1 つの文書とみなした．

$$D_{xl} = \bigcup_{i=100(l-1)+1}^{100l} T_{xi} \left(1 \le l \le 4\right) \quad (1)$$

例えば，D_{Adam1}は Adam の 1-100 件目のつぶやきの集合である．

ベクトルの要素は，履歴書とつぶやきに含まれる名詞とその複合語である．名詞と複合語の抽出には，それぞれ形態素解析器 Mecab[27]と専門用語自動抽出システム TermExtract[28]を用いた．

TFIDF は，TF（Term Frequency:単語の出現頻度）と IDF（Inverse Document Frequency : 逆文書頻度）を乗じた重みである．文書d_iにおける単語 j の TFIDF 値w_{ij}は以下の式で表される．

$$w_{ij} = tf_{ij} \times idf_j = \frac{n_{ij}}{\sum_k n_{ik}} \times \log \frac{N}{N_j} \quad (2)$$

n_{ij}とは，文書d_iにおける単語 j の出現回数，N は参照文書の総数，N_jは単語 j を含む文書数である．この方式では，N には Web ページの総数[29]を，N_jには Yahoo!検索で得られた Web ページの件数を利用した．

類似度$sim(D_{xl}, R_u)(x \in \{Adam, Bob, Charlie, Dave, Ellen\}, u \in \{Adam, Bob\})$は以下のように計算できる．

$$sim(D_{xl}, R_u) = \cos(D_{xl}, R_u) = \frac{D_{xl} \cdot R_u}{\|D_{xl}\| \|R_u\|} = \frac{\sum_{j=1}^{m} w_{D_{xl}j} w_{Rj}}{\sqrt{\sum_{j=1}^{m} w_{D_{xl}j}^2} \sqrt{\sum_{j=1}^{m} w_{Rj}^2}} \quad (3)$$

5.2 評価

Adam の履歴書と全てのつぶやきの集合との類似度を図 1 に示す．また，図 2 は Bob の履歴書と全てのつぶやきの集合との類似度を表したものである．図の横軸は$D_{xl}(x \in \{Adam, Bob, Charlie, Dave, Ellen\}, l \in \{1,2,3,4,\{1,2,3,4\}\})$，縦軸は類似度を表している．

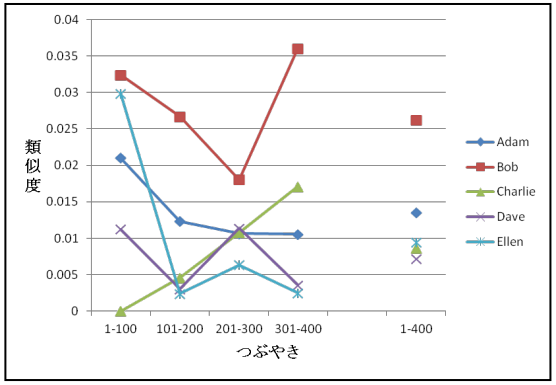

図 1　Adam の履歴書との類似度

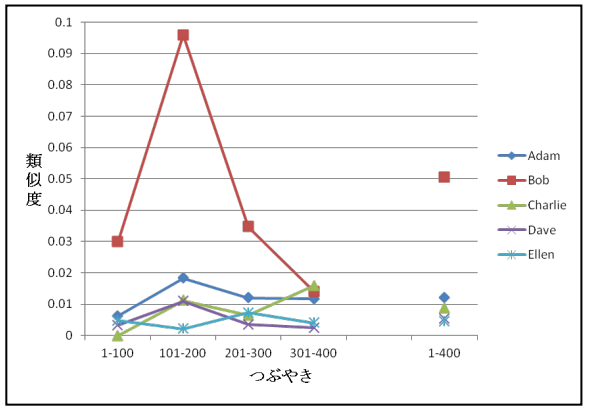

図 2　Bob の履歴書との類似度

図 1 で本人である Adam のつぶやきの類似度が最高となったのは 4 個の区分のうち 0 個，図 2 で Bob（本人）のつぶやきの類似度が最高となったのは 4 個の区分のうち 3 個であった．また，400 件のつぶやきを 1 つの文書とみなした場合，Bob の履歴書では Bob の類似度が最も高くなるが，Adam の履歴書では Adam の類似度よりも Bob の類似度の方が高くなる．このことから，従来手法を利用したベクトル空間モデルでは，例え本人のつぶやきであっても，自身の履歴書との類似度が常に最高となることが少ないことが分かる．

5.3 考察

他人のつぶやきの類似度が本人のつぶやきの類似度よりも高くなる原因について考察する．大きな原因として，履歴書に含まれる語句がつぶやきの中で直接使われる機会が多くないことが挙げられる．たとえば「電気通信大学」の学生は，電気通信大学という公式名称を用いるかわりに，電通大や UEC といった略語を利用することが多い．また，履歴書中の語句は個人情報につながるものが多いため，プライバシーを強く意識する人は，「調布の大学」などのように，特定のコミュニティに属する人の間のみで通じる言葉に言い換えることもある．このように直接記載されない語句については，つぶやきで示唆されていたとしても，TFIDF のように語句の出現頻度を利用する手法を用いた場合，類似度に反映させることができない．

図 3 は，Adam のつぶやき（101-200 件）から作成されるベクトルの重みを表したものである．横軸は Adam の履歴書に含まれる語句のうち TFIDF 値が高い 30 の語句を表す．これらの語句は類似度の寄与に大きく影響する．また，縦軸はこれらの語句の式(2)の値をベクトルの重みとして表している．

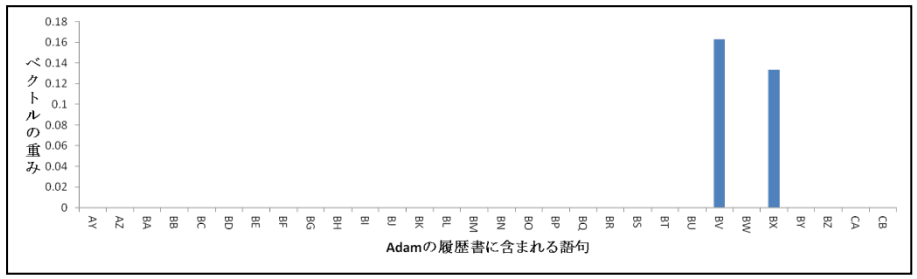

図 3　Adam のつぶやきから作成されるベクトルの重み

図から分かるように，履歴書の著者のつぶやきであるにも関わらず，履歴書のほとんどの語句の TFIDF 値が，つぶやきでは 0 となる．そのため，履歴書とつぶやきの類似度が不当に低くなってしまう．

6. WBC Advanced

6.1 方式

履歴書内の語句がつぶやき内で直接使われないため，出現頻度による類似度算出ができないという問題を回避するためには，履歴書に含まれる語句がつぶやき内で示唆されているときに，それを類似度に反映する必要がある．そのために，履歴書中の語句 j のつぶやき内での示唆の強さを$HTF(j)$とし，それを自動的に定量化する手法を検討する．HTF は Hidden Term Frequency（隠れ出現頻度）の略である．以下に，その詳細を示す．

I. 文章から名詞とその複合語を形態素解析により抽出する．抽出した語句の個数を n とする．

II. その中から最大 m 件の語句の組合せをクエリーとして抽出し，クエリーリストを作成する．クエリーの総数 N は以下の式で表される．

$$N = \sum_{i=1}^{m} \binom{n}{i} \tag{4}$$

III. クエリーリストに含まれるクエリー$q_i(1 \le i \le N)$について，以下の処理を繰り返し行う．

A) クエリーq_iを Web 検索し，検索結果の上位 k 件のタイトルを取得する．

B) $l(1 \le l \le k)$番目のタイトルについて，履歴書の語句 j が含まれているか探索し，以下の式によりf_lの値を返す．

$$f_l = \begin{cases} 1 & (j\text{が含まれている場合}) \\ 0 & (j\text{が含まれていない場合}) \end{cases} \tag{5}$$

C) f_lを用いてq_iのスコア$score(q_i)$を以下のように計算する．この値は 0 から 1 の値に正規化されている．

$$score(q_i) = \frac{1}{C}\sum_{l=1}^{k}(k-l+1)f_l \tag{6}$$

なお，C は以下の式で表される値である．

$$C = \sum_{l=1}^{k}(k-l+1) \tag{7}$$

IV. 全てのクエリーq_iの重み$score(q_i)$から，その最大値を選定し，$HTF(j)$とする．

$$HTF(j) = \max_i(score(q_i)) \tag{8}$$

得られた値は，人が履歴書内の語句とつぶやきとの関連性を判断する指標を定量化した値となる．今回は，m = 3，k = 20 とする．例えば，検索結果 20 件全てに履歴書の単語が含まれていたら HTF の値は 1 となり，19 件目と 20 件目に含まれていたら，値は$(2+1)/210 \cong 0.014$となる．提案手法を用いることで，例えば，「大学院入試説明会が今週 5 月 22 日土曜日に開催されます．同時に研究室公開も実施します」というつぶやきと「電気通信大学」との間に関連性があることを調べることが可能となる．なお，この場合の HTF の値は 0.14 となる．

次に，HTF をベクトル空間モデルに適用する方式を検討する．i 番目のつぶやきにおける単語 j の HTF を$h(i,j)$とおくとき，式(3)における文書ベクトルD_{xl}における単語 j の重み$w_{D_{xl}j}$を以下のように変更する．n はD_{xl}に含まれるつぶやきの件数である．今回は，n = 100 である．

$$w_{D_{xl}j} = \frac{1}{n}\sum_{i=1}^{n} h(i,j) \times idf_j \tag{9}$$

履歴書のベクトルの重み付けは予備実験と同様に，TFIDF を用いた．また，類似度計算には余弦を用いた．また，ベクトルの要素数は履歴書に含まれる全ての名詞数となる．これは，履歴書に含まれる名詞以外の HTF を求める必要がないからである．

6.2 評価

履歴書は前回の実験と同じサンプルデータを用いた．また，つぶやきの区切り方は前回の実験と同様である．以上の条件で実験を行い，提案方式の評価を行った．図 4 は Adam の履歴書との類似度を，図 5 は Bob の履歴書との類似度を表している．

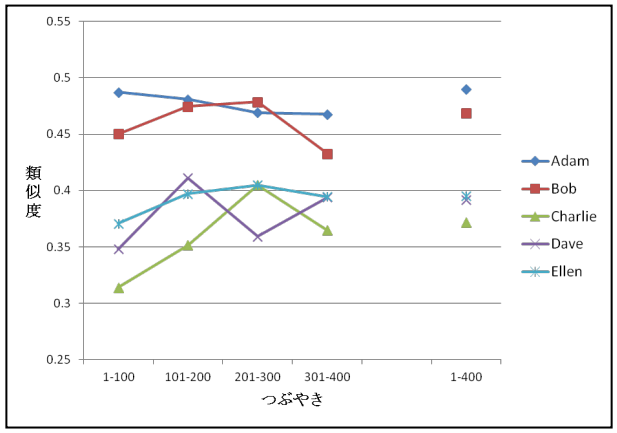

図 4　Adam の履歴書との類似度

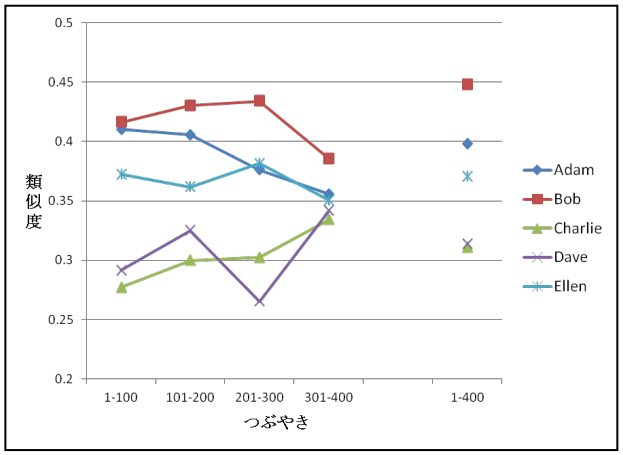

図 5　Bob の履歴書との類似度

図 4 では，本人である Adam のつぶやきの類似度が最も高かったのは 4 個の区分のうち 3 個，図 5 では Bob（本人）のつぶやきの類似度が最も高かったのは 4 個の区分のうち 4 個であった．また，400 件のつぶやきを 1 つの文書にまとめた場合，本人のつぶやきの類似度が最も高くなっていることも分かる．

6.3 考察

前回の手法と提案手法を利用した場合に作成されるベクトルの重みを比較する．図 6 は，図 3 と同じ語句について提案手法における式(8)の値をベクトルの重みとして追加したものである．

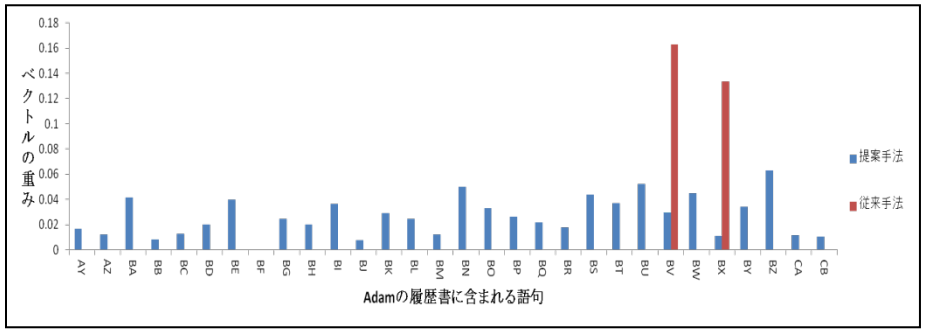

図 6　Adam のつぶやきから作成されるベクトルの重み

従来手法では，履歴書内のほとんどの語句がつぶやきに直接現れないため，類似度の算出に寄与していなかった．提案手法を用いることで，つぶやき内に直接記載されていなくてもその語句がつぶやき内で示唆される強さが定量化され，類似度の算出に寄与していることが分かる．

6.4 追加評価

提案手法の有効性を確認するために，サンプルデータに Twitter アカウントを追加して評価を行った．追加した 5 名のユーザの属性はそれぞれ表 2 の通りである．

Vol.2011-CSEC-55 No.2
2011/12/5

表 2　ユーザの属性

ユーザ名 (仮名)	出身大学	職場	専門分野
Frank	国立総合 F 大学	A 大学（准教授）	インタフェース
George	-	スポーツ選手	スポーツ
Henry	G 高専	H 工業系企業	機械工
Ivan	I 高専	Web 系 J 社	Web プログラミング
Justin	-	国立総合 K 大学生	文学

また，評価結果を図 7 と図 8 に示す．図 7 は Adam の履歴書との類似度を，図 8 は Bob の履歴書との類似度を表している．

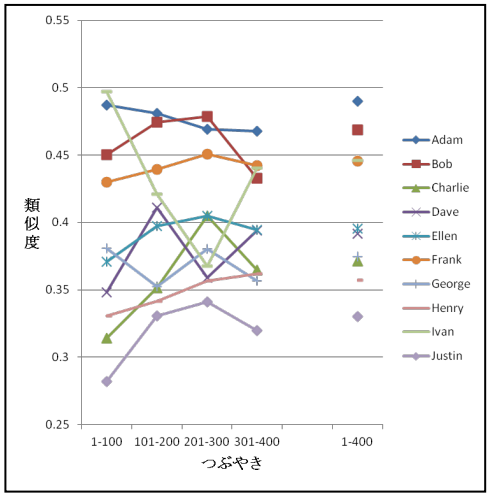

図 7　Adam の履歴書との類似度

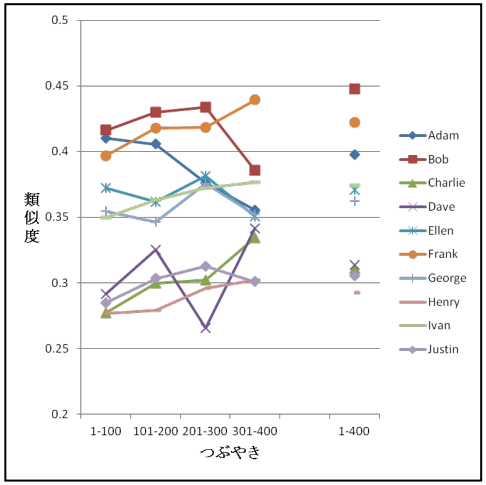

図 8　Bob の履歴書との類似度

Adam の履歴書では，Ivan のつぶやきの類似度が Adam のものよりも高くなることがある．これは，Ivan の出身大学，職場，専門分野が全体的に Adam に類似しているからであると考えられる．しかし，400 件を 1 つの文書とみなした場合は，Adam の類似度が一番高くなり，かつ他人と差が生じることが分かる．また，Bob の履歴書では，Bob と職場が同じである Frank のつぶやきの類似度が Bob のものより高くなることがある．しかし，同様に 400 件を 1 つの文書とみなした場合は，Bob の類似度が一番高くなり，かつ他人との差が生じる．このことから，文書に含むつぶやきの件数を増やすことで，安定して類似度を算出できると考えられる．

7. まとめと今後の課題

同一人物に関する複数の個人情報を第三者が入手し統合した場合，それぞれ単独の情報からでは分からなかった情報が明らかになる可能性がある．この問題の危険性を明らかにするために，Background Check とよばれる採用希望者への身辺調査を例に挙げ，ソーシャルメディアのプロフィールが匿名化されていても，履歴書の情報を基にソーシャルメディアのコンテンツの特徴を分析することで，採用希望者のアカウントを特定できることを示した．これにより，履歴書の情報とソーシャルメディアで開示された情報を統合し，個人の言動を調査することができる．この履歴書とソーシャル

メディアの照合の精度を向上させるために，コンテンツに含まれる語句と履歴書との間接的な関わりを定量化する手法を提案した．具体的な履歴書と Twitter のつぶやきを用いた評価を行い，提案手法の有効性を明らかにした．今後の課題は以下の通りである．

1. 履歴書の数と Twitter アカウントを増加させた大規模評価
2. 提案手法による攻撃からコンテンツの匿名性を守る手法の提案

参考文献

1) Lawler, J. P. and Molluzzo, J. C.: A Study of the Perceptions of Students on Privacy and Security on Social Networking Sites (SNS) on the Internet, Journal of Information Systems Applied Research, Vol.3, No.12, pp.1-18 (2010)
2) THINK SOCIAL「mixi の「いま」がわかるインフォグラフィック」 http://pr.mixi.co.jp/2011/06/01/infographics.html
3) livedoor ニュース「「レイプ容認」発言で炎上立教大生　内定先大手百貨店に「電突」騒ぎ」 http://news.livedoor.com/article/detail/5359699/
4) livedoor ニュース「【Sports Watch】店員女性がハーフナー・マイクにツイッターで悪口雑言、炎上騒ぎに」 http://news.livedoor.com/article/detail/5569059/
5) 電子情報通信学会: セキュリティハンドブック, オーム社 (2004)
6) Hacigumus H. et al.: Executing SQL over Encrypted Data in the Database-Services-Provider Model, In Proc. of the 2002 ACM SIGMOD International Conference on Management of Data, pp.216-227 (2002)
7) 志村正法, 宮崎邦彦, 西出隆志, 吉浦裕: 秘密分散データベースの構造演算を可能にするマルチパーティプロトコルを用いた関係代数演算, 情報処理, Vol.51, No.9 pp.1563-1578, 2010
8) Sweeney L.: k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol.10, No.5, pp.557-570 (2002)
9) Ashwin, M. et al.: V.:1-diversity: Privacy beyond k-anonymity, In Proc. of the 22nd IEEE International Conference on Date Engineering (2006)
10) Ninghui, L. et al.: t-closeness: Privacy beyond k-anonymity and l-diversity, In Proc. of the 23rd International Conference on Data Engineering, pp.106-115 (2007)
11) Krumm, J.: Inference Attacks on Location Tracks, In Proc. of the Fifth International Conference on Pervasive Computing, pp.127-143 (2007)
12) Bettini, C. et al.: Protecting privacy against location-based personal identification, In 2nd VLDB Workshop on Secure Data Management, pp.185-199 (2005)
13) Beresford, A. and Stajano, F.: Location privacy in pervasive computing, IEEE Pervasive Computing, Vol.2, No.1, pp. 46-55 (2003)
14) Acquisti, A. and Gross, R.: Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook, In Workshop on Privacy Enhancing Technologies, pp.36-58 (2006)
15) Meeder B., et al.: RT@ IWantPrivacy: widespread violation of privacy settings in the Twitter social network, In Proc. of the Web 2.0 Privacy and Security Workshop (2010)
16) Hart, M., et al.: Usable Privacy Controls for Blogs, In Proc. of the International Conference on Computational Science and Engineering, pp.401-408 (2009)
17) 渡辺夏樹, 片岡春乃, 内海彰, 吉浦裕: SNS 上のテキストからプライバシー情報を検知するシステムの構想と予備評価, 日本セキュリティマネジメント学会誌, Vol.24, No.3, pp.15-30 (2011)
18) Narayanan, A. and Shmatikov, V.: Robust De-anonymization of Large Sparse Datasets, In Proc. of the 29th IEEE Symposium on Security and Privacy, pp.111-125 (2008)
19) Narayanan, A. and Shmatikov, V.: De-anonymizing Social Networks, In Proc. of the 30th IEEE Symposium on Security and Privacy, pp.173-187 (2009)
20) Backstrom, R. et al.: Wherefore art thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography, In Proc. of the 16th International World Wide Web Conference, pp.181-190 (2007)
21) Novak, J. et al.: Anti-Aliasing on the Web, In Proc. of the 13th international conference on World Wide Web, pp.30-39 (2004)
22) Johnson v. K Mart Corp., No. 1-98-2172, http://www.state.il.us/court/opinions/appellatecourt/2000/1stdistrict/january/html/1982172.htm
23) Roberts, S. and Clark, L.: Myspace, Facebook, and Other Social Networking Siters: How Are They Used by Human Resource Personnel ?, In Delta Pi Epsilon National Conference, pp.35-43 (2008)
24) Greenwald J.: Web-based screening may lead to bias suits, Business Insurance, Vol.42 (2008)
25) Salton, G. et al.: A vector space model for automatic indexing, Communications of the ACM, Vol.18, No.11, pp. 613-620 (1975)
26) Jones, K.: A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, Vol.28, No.1, pp.11-21 (1972)
27) T. Kudo: MeCab: Yet another part-of-speech and morphological analyzer, http://mecab.sourceforge.net/
28) 専門用語（キーワード）自動抽出システム, http://gensen.dl.itc.u-tokyo.ac.jp/
29) The official Google Blog: we knew the web was big…, http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html