Webからの転移学習を用いたSNS検索

片岡　大祐†　　田島　敬史†

† 京都大学大学院情報学研究科　〒606–8301 京都府京都市左京区吉田本町
E-mail: †{kataoka@dl.soc.i.kyoto-u.ac.jp, tajima@i.kyoto-u.ac.jp}

あらまし　本論文では，SNS検索において，Web検索によって得られたデータセットからユーザ属性を識別するのに有用な語彙を学習させることで，投稿者の属性を指定した検索・ランキングを可能とする手法を提案する．単純に投稿者の属性をクエリに追加したSNSの投稿に対するキーワード検索は，投稿自身がショートテキストであり，その中に情報発信者のプロファイル情報が表現されていないため，検索性能を向上させることが難しい．そのため，間接的に学生であることを推定できるような語彙の情報が必要となるが，文書の特性上，SNSと比較してWeb検索の方がそのような語彙が得られやすいと考えられる．また，Twitterで学習用のデータを取得する場合，APIの制約やプロファイルやツイートなど様々な形態の文書からなるため，容易にデータを取得することが難しいという問題点がある．そこで，本研究では，SNSユーザ属性を推定するために必要な語彙を得るため,「学生」に対する「社会人」のように，相対する属性名を用意する．属性名をクエリとしたWeb検索で正例・負例を学習させる．その上で，投稿文と各投稿に関連づけられたプロファイルや過去の投稿などの文脈から，投稿の適合性を判定し，検索順位を決定する．

キーワード　ショートテキスト　転移学習　クエリ拡張

1. はじめに

近年，Facebook(注1)やInstagram(注2)などのソーシャルネットワーキングサービス（SNS）の普及により，誰もが情報発信や情報共有を気軽にできるようになった．その中でも，Twitter(注3)は代表的なSNSの1つであり，3億人を超えるアクティブユーザ数を有している．

Twitterは情報拡散が早いという特徴があるため，企業はイベントや新商品のプロモーションをTwitter上で行うことがよくある．さらに，実際にそのイベントに参加した人や新商品を購入した人の生の声をTwitter上で閲覧することができるため，Twitterはマーケティングにも使用されている．一方，同じ商品やコンテンツに対する評判が年代や性別，職業などのプロファイルによって異なるため，ユーザ属性ごとのカテゴライズも同様に需要がある．例えば，新しく発売されたiPhone Xが，若者は「使いやすい」と好評だったとしても，年配者には「使いにくい」という感想が多い場合がある．そこで，本論文では，投稿者の属性を指定した検索・ランキングを可能とする手法を提案する．

しかし，既存のTwitterの検索では以下の問題点が存在する．

P1:　投稿文がショートテキストのため，キーワードを追加した場合，適合率は増加するが，再現率が通常のWeb検索よりも低くなってしまう

P2:　投稿者の性別や年代，所属などの情報が投稿文中に記述されることは稀であるため，属性を表すキーワードを追加しても適合率・再現率の向上が見込まれにくい

P2の具体例を挙げると，例えば自分が学生や社会人であるといった情報を各投稿中に記載することは少ない．また，ソフトウェアエンジニアや京都大学といった具体的な所属名と比べ，学生や社会人といった情報はプロファイル中に記述されにくいため，プロファイルの検索から得ることも難しい．

本論文では，推定対象の属性名をクエリとし，その属性に相対する属性をセットで用意する．さらに属性名をクエリとしたWeb検索によって得られた検索結果を正例と見なし，相対する属性に対しても同様の操作を行い，得られたデータセットをもとに学習を行う．そのようにして得られた分類器をもとに，Twitterの投稿文とその周辺情報を利用して，投稿文中からの判断が難しい属性を推定することを考える．また，推定対象の属性名をクエリとした検索結果を正例とするため，負例には同粒度の対象的な属性名をクエリとした検索結果を与える．例えば，推定したい属性が「学生」だった場合，Web検索に「学生」というクエリを投入し得られた検索結果を正例とみなし，相対する「社会人」というクエリに対しても同様の操作を行い，得られた検索結果を負例とする．

ここで，データセット作成にSNS自体の検索ではなく，Web検索を用いる理由として，以下の仮説を立てた．

H:　文書の傾向として，話し言葉で書かれた投稿が多いSNSと比べ，Webの文書は書き言葉が多く，属性推定に必要な語彙を得られやすい．

SNSで「学生」というクエリで検索を行った場合，検索結果の文書中に共起する語には「学生」と関連する語である可能性が高いと考えられる．しかし，Web検索で同様の操作を行った場合,「学生」を説明するような形式の文書が多いため，SNSでの同様の操作では得られないような，学生を推定するのにより多くの有用な語彙が得られると考えた．

このようにして得られた分類器から，推定したい属性に合致

(注1)：https://www.facebook.com
(注2)：https://www.instagram.com
(注3)：https://twitter.com

するかどうかの推測値を算出する．分類器の入力文書としては，任意のトピックをクエリとし，得られた検索結果とその投稿者のプロファイルや過去のツイート，フォロワー・フォロイーのプロファイルを 1 つの文書とみなした投稿者を特徴づける文書ベクトルを用いる．推測値をもとにリランキングされた検索結果を出力する．

提案手法の有効性を評価するために，10 種類のトピックを表すクエリと属性を表すクエリのペアからなるテストコレクションを作成した．Twitter 検索，Web 検索それぞれにおいて属性クエリを入力した検索結果からデータセットを作成し，学習を行なった．Twitter の学習と比べ，Web の学習の結果，平均 nDCG@10 が 0.338 から 0.382 に向上した．

我々の主な貢献を以下にまとめる．1)SNS でユーザ属性を考慮した検索を行うための手法を提案した．2)Web 上で学習した対象ユーザ属性の語彙を用いて，投稿者の周辺情報からユーザ属性に該当するかを判定する手法を提案した．3) テストコレクション，および，提案手法の評価の枠組みを構築し，提案手法の有用性を示した．

本論文の構成は以下のとおりである．2 節では，関連研究を紹介する．3 節では，本研究で扱う問題の定義を行う．4 節では，提案手法について述べる．5 節では，提案手法に対する実験と評価について述べる．6 節では，今後の課題及び本論文での結論について述べる．

2. 関 連 研 究

本節では，本研究と関連する研究について言及し，本研究の位置づけについて述べる．

SNS をターゲットドメインとして転移学習を利用した研究が行われている．Peddinti ら [1] は，映画のレビューから Twitter ドメインに EM アルゴリズムと Rocchio SVM を用いた Domain Adaptation を適用することで Twitter の投稿の感情分析を行った．Lu ら [2] は，メンション中の抽象的な表現の属性推定を行った．Lu らの手法では，語彙の不足を補うために，ユーザ間の他のメンションや Wikipedia の語彙を利用している．

適合フィードバック･擬似適合フィードバックを用いて SNS での検索性能の向上に取り組んだ研究について述べる．Miyanishi ら [3] は，検索者が検索結果のマイクロブログ文書の中から 1 つ正解文書を選択し，その文書をクエリ拡張に用いて再検索した検索に対し擬似適合フィードバックを適用することで検索精度の向上を図った．Whiting ら [4] はマイクロブログの文書に対し，PageRank を用いて文書中の単語に時間性を考慮した重み付けを行い，時間性を考慮した疑似適合フィードバックを行うことで検索性能が向上することを示した．我々の以前の研究 [5] では，本研究と同様の問題設定に取り組んだ際に，検索者自身が投稿者の周辺情報を確認することで，正例・負例を付与する適合フィードバックを導入した．

また，本研究では，投稿文のみの語彙を使用するのではなく，投稿文からリンクで辿ることの出来る投稿者のプロファイルや過去のツイート，フォロワーのプロファイルなどから属性を推定する．同様にリンクで関連づけられた情報を用いて属性推定を行った研究について述べる．奥谷ら [6] は，Twitter において，ユーザ間のやりとりであるメンション情報を利用して投稿者のプロファイル推定を行った．上里ら [7] は，推定対象となるユーザの相互フォロー，相互メンション関係にある周辺ユーザの属性情報を補完することで属性推定の精度を向上させた．また，上里らは，上記の研究を発展させ，Personalized PageRank を用いて推定対象ユーザと周辺ユーザの関連度を算出し，関連度に応じて単語の重みを変化させることで属性推定の精度を向上させる手法を提案した [8]．池田ら [9] は，ツイート投稿者の普段のツイートの傾向を解析することにより，プロファイルの推定を行った．

Web 上のデータの検索においても異なるデータセット間での検索に関する研究が行われている．Halpin ら [10] は Semantic Web の検索とハイパーテキストの Web データの検索において一方の検索結果のフィードバックを他方の検索に用いることで検索精度を向上させることができることを示した．Herzig ら [11] は，1 つのデータセットの語彙のみを用いてクエリを拡張することで，RDF で記述された Web 上の異なるデータセット間の検索を行う手法を提案した．Shekarpour ら [12] は，検索者が与えたキーワードを適切に解釈するために，予め定義された基本グラフパターンテンプレートを用いて，SPARQL クエリを生成する手法を提案した．Su ら [13] は，自然言語で記述されたクエリをグラフデータに変換し，Knowledge Graph での検索により得られたデータに対し適合フィードバックを行うことで検索性能が向上することを示した．

同様の問題設定に取り組むという点で，宮西ら [3] の研究，我々の以前の研究 [5] に類似している．これらの研究では，検索者が明示的なフィードバックを行うことで検索性能を向上させているが，インタラクションが増えるという点で検索者への負担となりうる．本研究では，この問題を解決するアプローチとして，Web 検索によって得られたデータセットを用いて学習を行い，投稿文とその周辺情報から投稿者が対象とするユーザ像に合致するかを判定し，検索結果のリランキングを行う．

3. 問 題 定 義

本節では，我々が取り組む問題の詳細について述べる．

3.1 問 題 設 定

本節では，本研究が取り組む問題設定について述べる．本研究では，**トピッククエリ**と**属性クエリ**の 2 つのクエリを用いる．トピッククエリは，検索対象の文書に対する問い合わせに用いるためのクエリであり，属性クエリは，Web におけるデータセットの収集と属性の指定のために用いる．例えば，新発売の商品 X に対する学生の評判を知りたいという検索欲求に対しては,「X」がトピッククエリ,「学生」が属性クエリである．この場合，X で検索し得られた検索結果に対し「学生」という属性クエリを入力することで，事前に学習したモデルを用いて「学生」による投稿である可能性が高い検索結果を上位にリランキングする．その際，投稿者のプロフィール，過去のツイートや，フォロワー・フォロイーのプロフィールから，各ユーザごとに生成した文書ベクトルを用いて，分類器にかけ，投稿者が「学

生」であるかの推定値を出力する．出力した推定値をもとに検索結果をリランキングする．

3.2 ベクトル空間

本研究の検索を実現するにあたって，文書の表現にはベクトル空間モデルを用いる．本研究で扱うコーパスは，Twitter，Web それぞれに出現する語を合わせたものからなる．コーパスの全ての語集合を V とすると，ある文書 d の特徴ベクトル $\mathbf{x}$ の次元数は $|V|$ であり，$\mathbf{x} \in \mathbb{R}^{|V|}$ である．

3.3 転移学習の設定

本節では，我々が行う転移学習のシナリオについて述べる．神嶌ら [14] の分類に基づくと，本研究での転移学習の位置付けは，ラベルのあるソースドメインから，ラベルのないターゲットドメインへの知識の転移を行うトランスダクティブ転移学習である．より正確には，Web 検索で得られた検索結果は対象の属性を推定するために有用な語彙を含む正例の文書であると見なして正例を作成する．

3.4 転移学習の定義

本節では，ユーザ属性を考慮した SNS 検索のための転移学習の定義を与える．ドメインは，定義域 $\mathcal{X}$ と周辺確率 $P(X)$ によって特徴付けられ，$\mathcal{D} = (\mathcal{X}, P(X))$ と表される [15]． また，ラベル集合 Y と，ドメイン $\mathcal{D}$ から $P(X)$ に従い得られたインスタンス集合 $X \subset \mathcal{X}$ とのペアを，ドメインデータと呼び，$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots\}$ で表す．本研究において，ソースドメインは Web であり，$\mathcal{D}^{(S)}$ と表され，ターゲットドメインの Twitter は，$\mathcal{D}^{(T)}$ と表される．データセット作成のためのクエリは，ラベル集合 Y の要素 $q \in Y$ で表現される．また，5.2.1 節にて詳細は後述するが，各インスタンス $\mathbf{x}$ は，検索結果の文書 d_0 を含む複数の文書から構成されており，検索対象データ n 件取得する場合，$R = \{d_{0,1}, d_{0,2}, \ldots d_{0,n}\}$ と表される．システムはランクづけられた文書集合 R を出力する．ランクはクエリ q および，ターゲットドメイン $\mathcal{D}^{(T)}$ から得られたインスタンス集合 $X^{(T)}$ から推定されるランク関数 $f_q : X^{(T)} \to \mathbb{R}$ によって与えられる．すなわち，順序 $\succ$ が，$X^{(T)}$ 上に $d_i^{(T)} \succ d_j^{(T)} \Leftrightarrow f(\mathbf{x}_i) \succ f(\mathbf{x}_j)$ で定義される．また，得られる順序集合 $(R, \succ)$ はある検索結果の評価指標によって評価されるものとする．この時，本研究が取り組む転移学習は，この検索評価指標の値を最大化するランク関数 f を推定する問題であると定義できる．

4. 提 案 手 法

本節では，我々が提案する転移学習の手法について詳細を述べる．本研究の目的は，Web 検索から得られたデータセットをもとに学習を行い，SNS の任意のクエリを投入し得られた検索結果に対して属性を推定し，求めるユーザ像に合致する投稿を上位にリランキングすることである．

本手法の学習は，以下の 3 つのステップから構成される．

（1） 推定対象の属性集合を用意し，「学生」に対する「社会人」のように同系列の属性ごとにグループ分けをする

（2） 同系列のクエリに対してそれぞれ Web 検索を行う

（3） 「学生」を正例とした場合，「社会人」が負例のように相対する属性ごとに，データセットを作成し，学習を行う

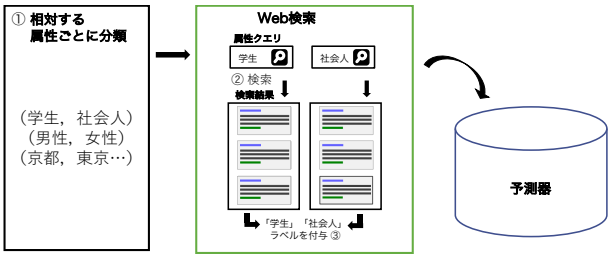

図 1 学習の概念図

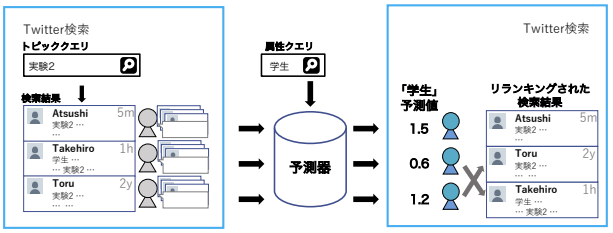

図 2 適用の概念図

学習の概念図を図 1 に示す．また，本手法の適用は，以下の 4 つのステップから構成される．

（1） トピッククエリ・属性クエリを入力

（2） トピッククエリを入力し得られた検索結果の各文書に対し，周辺情報も加えたベクトルを作成

（3） 学習により得られた回帰モデルにベクトルを入力として与え，属性クエリに適合するかどうかの予測値を出力

（4） 予測値の大きい順に検索結果をリランキング

検索への適用の概念図を図 2 に示す．

4.1 文書ベクトルの生成

本節では，Twitter で属性を予測するために用いる各検索結果の文書ベクトルの生成方法について述べる．本研究では，各ユーザの属性を推定するために，検索結果中のツイートに加え，投稿者のプロファイル，過去のツイート，フォロワーとフォロイーのプロファイルといった文脈から投稿者の属性を推定するのに有用な語彙を収集する．それらのツイートの文脈情報を用いて以前の我々の研究 [5] と同様の方法を用いてユーザベクトルを作成する．具体的には以下の通りである．まず，あるトピッククエリを Twitter の検索システムに入力し，n 件からなる検索結果のツイート列 $v_{0,1}, v_{0,2}, \ldots v_{0,n}$ が得られたとする．このとき i 番目の検索結果 $v_{0,i}$ に対し，ラベルつき有向グラフ $G = (V, E, L)$ を生成する．ノード $V = \{V_0, V_1, \ldots, V_k\}$ は $k+1$ 種類からなり，検索結果を除いて文脈中にノードが k 種類存在することを表す．ただし V_0 は $V_0 = \{v_{0,1}, v_{0,2}, \ldots v_{0,n}\}$ と表される検索結果の集合である．ノードは投稿者やフォロワーのプロファイル，過去のツイートが該当する．エッジは $E \subset V \times L \times V$ と定義され，$e = (v_{0,1}, l_1, v_{1,1})$ は 1 番目の検索結果のノード $v_{0,1}$ からノード $v_{1,1}$ に，l_1 というラベルのついたエッジが張られていることを表す．本研究でのノードは，検索結果からなる tweet ノード投稿者のプロファイルが付随する user ノード，フォロワーのプロファイルが付随する follower ノード，検索結果を除く他のツイートからなる user_tweet ノー

ドを用意する．

本研究でのエッジは次のように定義した．user ノードから tweet ノードに，post というラベルのついたエッジを張る．次に，ある投稿者が別の投稿者をフォローしている場合，follower ノードから user ノードに，follow というラベルのついたエッジを張る．次に，投稿者は検索結果のツイート以外にも投稿している場合，user ノードから user_tweet に，nearby-post というラベルのついたエッジを張る．最後に，検索結果中の投稿を除き，同じ投稿者による複数の投稿がある場合，前後の投稿同士，つまり user_tweet から user_tweet に connect というラベルのついたエッジを張る．

ノード v からノード v' へのパス $P_{v\to v'}$ における伝播率の積 $\delta(P_{v\to v'})$ を以下の式で定義する．

$$P_{v\to v'} = (e_1, e_2, \ldots, e_n) \tag{1}$$

$$\delta(P_{v\to v'}) = \prod_{i=1}^{n} d(l_{e_i}) \tag{2}$$

ただし，$(e_1, e_2, \ldots, e_n)$ は v から v' に到達するまでに通るエッジを順に並べたタプル，l_{e_i} はエッジ e_i に付随するラベル，n はノード v からノード v' に到達するまでに通るエッジの数である．$d(l_{e_i})$ はラベルの種類ごとに予め定められる伝播率であり，$0 \leqq d(l_{e_i}) \leqq 1$ を満たす．

そのため，i 番目の検索結果 $v_{0,i}$ から張られる SNS グラフ中の全ての文書ベクトル $\{\mathbf{x}_{0,i}, \mathbf{x}_{1,i}, \ldots, \mathbf{x}_{|V|,i}\}$ と予め設定した伝播率の積 δ，パラメータ α を用いて i 番目の検索結果の文書ベクトル $\mathbf{x}'_{0,i}$ は以下の式で表される．

$$\mathbf{x}'_{0,i} = (1-\alpha)\mathbf{x}_{0,i} + \alpha \sum_{V_j \in V}^{|V|} \sum_{v_{j,i} \in V_j \setminus v_{0,i}}^{|V|} \delta(P_{v_{j,i}\to v_{0,i}})\mathbf{x}_{j,i} \tag{3}$$

また，ある文書 d におけるある単語 t の重み $w(d,t)$ は TF-IDF スコアを用いて以下のように計算する．

$$w(d,t) = \mathrm{tf}_{d,t} \cdot (\mathrm{idf}_t + 1) \tag{4}$$

ここでの TF スコア $\mathrm{tf}_{d,t}$ は d における単語 t の出現回数とし，IDF スコア idf_t はそれぞれ以下のように計算する．

$$\mathrm{idf}_t = \log_2 \frac{N+1}{n_t+1} \tag{5}$$

但し，N は文書ベクトルの総数，n_t は単語 t が含まれる文書の数である．Twitter においてはノードを 1 文書とし，Web においては 1 つの検索結果を 1 文書とした．このようにして求められた各単語の重みをもとに，ユーザごとに文書ベクトルを生成する．

5. 実　　験

本節では，提案手法を用いて行った実験とその評価について述べ，結果についての考察を行う．本実験の目的は提案手法の有効性について検証することである．

5.1 概　　要

本研究では，予備実験，本実験の 2 つの評価実験を行なった．

予備実験の目的は，学習において最適な文書数を調べることである．これは，学習結果が入力する文書の数に依存すると考えたためである．初めに Twitter における属性の分類タスクを設定し，文書数を変化させた上で平均精度が高くなる文書数を確認する．本実験では，Twitter，Web それぞれに対し学習用データセットを作成するため，Twitter，Web それぞれの最適な文章数を確かめる．

本実験では，予備実験で得られたパラメータをもとに文書のスコアを導出するための回帰モデルを作成する．そのモデルを用いて，Twitter での属性クエリ・トピッククエリによる検索のランキングを行い，評価指標を用いて，検索性能の差を調べる．この際，以下の 3 手法を比較手法とした．

(1) Twitter 上で文脈情報を考慮した適合フィードバック [5]
(2) Twitter 上で学習させたモデルの Twitter 検索への適用
(3) Web で学習させたモデルの Twitter 検索への適用 (提案手法)

5.2 実 験 設 定

5.2.1 Twitter の文書ベクトルの作成

ユーザ属性を推定させるために節に従い，各ユーザに対し，ユーザとフォロワーのプロファイル，過去のツイートから文書ベクトルを作成した．単語の重みの調整は，式において，$\alpha = 0.75$ とし，エッジの伝播率に関しては，$d(\text{post}) = 1.0$, $d(\text{follow}) = 1.0$, $d(\text{nearby-post}) = 0.5$, $d(\text{connect}) = 0.5$ とした．これは以前の研究 [5] で最も良い精度となったパラメータと同様のものである．

5.2.2 学　　習

本研究で使用した機械学習アルゴリズムは，アンサンブル学習の一種である Random Forest [16] を使用した．Random Forest は，ブートストラップサンプルから複数の決定木で学習を行い，それぞれの決定木による分類結果を多数決により統合し最終的な分類結果とする．予備実験の 2 クラス分類においては，scikit-learn(注4) の Random Forest Classifier ライブラリを使用した．また．本実験においては，ランキングのために Random Forest Regressor ライブラリを使用した．Random Forest 回帰では，予測される目的変数を全ての決定木の予測を平均することで計算する．

5.2.3 学習データ

属性クエリは「学生」と「社会人」,「男性」と「女性」の 4 つを用意した．Web の学習データの取得には，Bing Web Search API(注5)を利用した．各属性クエリに対し，320 件の文書を取得した．Twitter の学習データの取得には，Twitter REST API(注6)を利用し，各クエリに対し，320 件のツイートを取得した．API の制限から，投稿者のフォロワー数が 1200 人以下の場合は全員，1200 人を超える場合 1200 人ずつフォロ

(注4)：http://scikit-learn.org/stable/
(注5)：https://azure.microsoft.com/ja-jp/services/cognitive-services/bing-web-search-api
(注6)：https://dev.twitter.com/rest/public

表 1　語彙の比較

	平均語彙	最大語彙	最小語彙
Twitter	2177.6	4607	62
Web	1147.9	14215	1

表 2　実験に使用したクエリ

	トピッククエリ	属性クエリ	正例の数	想定した検索意図
q_1	スマートスピーカー	学生	2	購入者の層を知りたい
q_2	スマートスピーカー	社会人	52	(同上)
q_3	今年の漢字	学生	19	興味のある層を知りたい
q_4	今年の漢字	社会人	42	(同上)
q_5	忘年会	学生	9	学生と社会人の考えの差を知りたい
q_6	忘年会	社会人	90	(同上)
q_7	モンハン	学生	42	ゲームがどの層に人気かを知りたい
q_8	モンハン	社会人	58	(同上)
q_9	甲子園	学生	25	どの層が言及しているかを知りたい
q_{10}	甲子園	社会人	35	(同上)
q_{11}	紅白	男性	29	男女別に人気のある歌手を知りたい
q_{12}	紅白	女性	23	(同上)
q_{13}	スターウォーズ	男性	59	男女の割合を知りたい
q_{14}	スターウォーズ	女性	35	(同上)
q_{15}	海月姫	男性	60	男女別人気を知りたい
q_{16}	海月姫	女性	29	(同上)
q_{17}	アベマ TV	男性	22	男女別人気を知りたい
q_{18}	アベマ TV	女性	65	(同上)
q_{19}	インフルエンザ	男性	34	男女の割合を知りたい
q_{20}	インフルエンザ	女性	46	(同上)

ワーの自己紹介文を取得した．投稿者の過去のツイートは検索結果のツイートの前後，それぞれ最大 100 件ずつ取得した．表 1 に学習に用いた Twitter と Web の 1 文書あたりの平均語彙，最大語彙，最小語彙を示す．Twitter のデータセットの特徴ベクトルは，文脈の語彙を含んでいるため，平均語彙は，Twitter の方が多い．しかし，Web の文書の方が 1 文書あたりの語彙の幅が大きいことが判明した．

5.2.4　評価データ

実験には，10 件のクエリを用いた．クエリを表 2 に示す．同様のトピッククエリに対して，複数の属性クエリの存在を認めた．各クエリに対して，100 件ずつツイートを取得し，取得した全てのツイートに対し，ツイート本文とプロファイル等を確認し，予め手動で適合性判定を行った．検索結果以外のツイートは各ユーザごとに前後最大 50 件ずつ取得し，フォロワーは API の制限から，最大 1200 件まで取得した．属性クエリと一致しているツイートを正例，属性クエリに一致していないツイートは負例，文脈情報を確認しても判定できなかった場合は判定不可として負例と同様の扱いとした．

5.2.5　単語の処理

各ノード中の文書の形態素解析には MeCab(注7)を利用した．また，特徴ベクトルの作成にあたって，文書から名詞のみを抽出した．頻度の少ない単語の影響を低くするため，コーパス中に 5 回以下しか出現しない単語を除去した．

5.2.6　評 価 指 標

本研究の目的は，トピッククエリと属性クエリの 2 つのクエリを用いて，文書を順付けることである．予備実験では，件数の異なる文書に対して分類を行うため，ROC 曲線の下面積である AUC を採用した．本実験では，検索手法の評価を行うため，評価指標として normalized Discounted Cumulative Gain(nDCG) を用いる．nDCG は多値適合性に基づく評価指標であり，検索順位を考慮した性能を比較できる．順位付けを k 位まで行う場合の nDCG は DCG@k [17] を理想的な順位で適合文書が並んだと仮定した DCG'@k で正規化した指標であり，DCG@k は以下の式で定義される．

$$\mathrm{DCG@}k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{6}$$

ただし，rel_i は i 番目の検索結果の関連度であり，今回の実験においては i 位の文書が適合文書の場合は $rel_i = 1$，不適合文書の場合は $rel_i = 0$ とした．

(注7)：https://code.google.com/p/mecab

表 3　年代・性別の分類結果の AUC

学習文書数	Twitter(年代)	Web(年代)	Twitter(年代)	Web(年代)
10	0.533	0.634	0.373	**0.568**
20	**0.546**	0.601	0.451	0.514
40	0.492	0.599	0.476	0.543
80	0.459	0.625	0.538	0.503
160	0.445	**0.636**	**0.577**	0.510
320	0.352	0.584	0.531	0.500

表 4　属性別の検索性能の平均

k	CRFG(年代)	Twitter(年代)	Web(年代)	CRFG(年代)	Twitter(年代)	Web(年代)
3	0.861	0.440	0.500	0.977	0.377	0.206
5	0.778	0.444	0.472	0.914	0.368	0.248
10	0.636	0.406	0.426	0.722	0.441	0.330
20	0.563	0.396	0.417	0.593	0.462	0.349
30	0.560	0.418	0.424	0.558	0.464	0.375

5.3　予 備 実 験

5.1 節で述べた予備実験で行う実験の内容とその結果，考察について述べる．予備実験では，以下の 2 通りの手法に対し，相対する属性の 2 クラス分類を行い，各属性クエリの平均精度が最大となる文書数を調べる．

(1) **Twitter:**　Twitter 検索で属性名をキーワードとした検索による分類モデルの学習

(2) **Web:**　Web 検索で属性名をキーワードとした検索による分類モデルの学習

文書数は，各属性ごとに $10, 20, 40, 80, 160, 320$ 件とし，2 つの属性を合わせて $20, 40, 80, 160, 320, 640$ 件からなる文書でそれぞれ学習させた 10 通りを比較する．分類する属性は，表 2 のクエリに示すように，学生/社会人の年代の分類，男性/女性の性別の分類の 2 通りである．年代・性別それぞれの分類タスクの実験結果は表 3 の通りである．表に記載している AUC はそれぞれ 5 件の分類タスクの平均 AUC である．

320 件まで学習文書数を増やした場合，Twitter と Web，年代と性別それぞれの組み合わせにおいて最も AUC の平均が高い学習文書数が異なるという結果になった．年代分類の Twitter の学習は 20 件，Web の学習は 160 件の文書数を用いた場合に，性別分類の Twitter の学習は 160 件，Web の学習は 10 件の文書数を用いた場合に最も平均 AUC が高くなった．本実験では，

上記の文書数を用いて，検索性能を調べる．

5.4 本 実 験

5.1 節で述べた本実験で行う実験の内容とその結果，考察について述べる．本実験では，予備実験で得られたパラメータを用いて，表 2 のトピッククエリと属性クエリを用いた検索を行う．本実験では，以下の 3 通りの手法を用いて検索性能を比較する．

(1) **CRFG:** 文脈の語彙を重みを調節して文書ベクトルに加えた Twitter 上での適合フィードバック [5]

(2) **Twitter:** 上記の (1) と同様

(3) **Web:** 上記の (2) と同様

CRFG では，上位 10 件をフィードバックとして与えてクエリを修正する．また，評価の際に，フィードバックに用いた 10 件のツイートは除外した．上記の 3 通りの手法を比較した結果の全体，属性別の平均を表 4 に，10 件のクエリの精度の推移を図**??**に示す．ただし，表 4 中の k は検索結果の上位 k 件に対して評価値を計算したことを表す．

実験の結果，適合フィードバックを用いる CRFG が全体平均においても各属性ごとの検索においても最も高い値を示した．一方，Twitter と Web を比較した場合，年代を属性クエリに指定した場合，Web の方が，性別を属性クエリに指定した場合，Twitter の方が，平均的に高い精度を示した．2 クラス分類では判定できなかった結果が，回帰モデルを用いたことで推定したい属性に該当するユーザを上位にランキングできていることが確認できた．

5.5 考 察

本節では，実験結果を踏まえた考察を行う．

予備実験において，学習文書数が増えても精度が向上しなかった原因であるが，Twitter では，投稿者が属性に該当する場合，投稿中に属性名が現れることが稀であるため，そもそも投稿者が属性クエリと一致しない場合が多いからだと考えられる．Web の場合は，性別の分類タスクの精度が低かったが，これは検索結果の下位に行くほど，両方の性に言及したノイズとなる文書が多くなっていたからだと考えられる．

本実験の結果，Twitter の学習の精度が低かった原因であるが，仮説通り，投稿内容が属性を言及対象としているからだと考えられる．一方，Web の性別の精度が低かったのは，予備実験で AUC が 0.5 付近であったように，男女の差を学習させることができていなかったことが挙げられる．しかし，学習に用いたベクトルの語彙の平均が Web より，Twitter の方が多いという設定で実験を行なっており，Web の方がより少ない語彙で学習できているといえる．また，Peddinti ら [1] のように，Domain Adaptation を用いて，分布を補正することで，Web の語彙を Twitter の語彙に適応させることも検討している．

6. おわりに

本論文では，属性名をクエリとした Web 検索を行い，得られた検索結果を用いて相対する属性を学習をすることで，Twitter においてトピックと投稿者の属性を指定した検索・ランキングを可能とする手法を提案した．

また，Twitter での検索性能を比較する実験の結果，Twitter での学習と比較し，Web での学習の方が高い精度を示した．

今後は，クエリ数と属性の数を増やし，より多くの文書を用いた学習を行う一方，有効な属性の発見や手法の拡張を行う予定である．

謝辞 本研究の一部は，JST，CREST（#JPMJCR16E3）によるものです．ここに記して謝意を表します．

文 献

[1] Viswa Mani Kiran Peddinti and Prakriti Chintalapoodi. Domain adaptation in sentiment analysis of twitter. *Analyzing Microtext*, 11:05, 2011.

[2] J-L Lu, M. P. Kato, T. Yamamoto, and K. Tanaka. Entity identification on microblogs by CRF model with adaptive dependency. *The Institute of Electronics, Information and Communication Engineers Transactions*, pages 2295–2305, 2016.

[3] T. Miyanishi, K. Seki, and K. Uehara. Improving pseudo-relevance feedback via tweet selection. In *Proceedings of the 22nd ACM international conference on Information and Knowledge Management*, pages 439–448, 2013.

[4] S. Whiting, I. A. Klampanos, and J. M. Jose. Temporal pseudo-relevance feedback in microblog retrieval. In *ECIR*, pages 522–526, 2012.

[5] D. Kataoka, M. P. Kato, T. Yamamoto, O. Hiroaki, and K. Tanaka. Context-aware relevance feedback over sns graph data. In *Proceedings of the International Conference on Web Intelligence*, pages 823–830. ACM, 2017.

[6] 奥谷貴史，山名早人. メンション情報を利用した twitter ユーザープロフィール推定. **日本データベース学会和文論文誌**, 13(1):1–6, 2014.

[7] 上里和也，浅井洋樹，奥野峻弥，山名早人. Twitter ユーザを対象とした属性推定の精度向上-周辺ユーザの属性補完を利用して-. In **第 7 回データ工学と情報マネジメントに関するフォーラム**, 2015.

[8] 上里和也，浅井洋樹，山名早人. Personalized pagerank を利用した網羅的 Twitter ユーザ属性推定. In **第 8 回データ工学と情報マネジメントに関するフォーラム**, 2016.

[9] 池田和史，服部元，松本一則，小野智弘，東野輝夫ほか. マーケット分析のための twitter 投稿者プロフィール推定手法. **情報処理学会論文誌コンシューマ・デバイス & システム** *(CDS)*, 2(1):82–93, 2012.

[10] H. Halpin and V. Lavrenko. Relevance feedback between web search and the semantic web. In *IJCAI*, pages 2250–2255, 2011.

[11] D. M. Herzig and T. Tran. Heterogeneous web data search using relevance-based on the fly data integration. In *WWW*, pages 141–150, 2012.

[12] S. Shekarpour, S. Auer, A. C. Ngonga Ngomo, D. Gerber, S. Hellmann, and C. Stadler. Generating sparql queries using templates. *An International Journal Web Intelligence and Agent Systems*, pages 283–295, 2013.

[13] Y. Su, S. Yang, H. Sun, M. Srivatsa, S. Kase, M. Vanni, and X. Yan. Exploiting relevance feedback in knowledge graph search. In *KDD*, pages 1135–1144, 2015.

[14] 神嶌敏弘 et al. 転移学習. **人工知能学会誌**, 25(4):572–580, 2010.

[15] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016.

[16] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[17] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM TOIS*, pages 422–446, 2002.