1K2-5

テキストマイニングを用いた転職サイトの会員離脱予測

Churn prediction by using text mining for job search sites

上門 雄也 *1 Kazuya Uekado　大和田 勇人 *2 Hayato Owada　金盛 克俊 *3 Katsutoshi Kanamori　鈴木 正昭 *4 Masaaki Suzuki

*1*2*3*4東京理科大学 理工学部 経営工学科
Department of Industrial Administration, Tokyo university of science

Companies managing a job search site earn revenue from posting a want ad. Clients want to post a want ad on job search sites that have many active users. Therefore, it is important for companies managing a job search site to increase active user. However, many users never apply after registering. This paper proposes a method to find out features of churn. In the previous research, we can't comprehend indwelling features because they have two problems. First, classification process is a black box. Second, they use only registration information and activity log as features. Our method can comprehend indwelling features by using text mining to resume of specialized work experience. As a result, users writing 'sales' and 'development' tend not to do leave. In this research, company can comprehend the site's strong points and weak points. Moreover, we can predict churn finely.

1. はじめに

1.1 背景

現在，日本の多くの企業では終身雇用制度を採用している．しかし，近年ではアメリカの企業にならい，成果主義的な人事制度を導入し始めた [1]．これにより，日本の終身雇用制度は崩壊しつつあるといわれ，転職をすることが一般的となった．

また，近年では急速に情報化が進み，多くの人がインターネットを利用している．インターネットの普及とともに，転職経路は大きく変化した．2002 年では「人的つながり」や「新聞広告・チラシ」が主流の転職経路で，「転職サイト」はわずか 4.8%の人しか利用していなかった．しかし，2010 年になると転職経路はインターネット中心となり，「転職サイト」を 62.0%の人が利用するまでに増加した [2]．

この転職サイトとは，転職希望者と求人企業をマッチングする目的や機能をもったサイトを意味する．転職サイトを運営する企業は，求人広告の掲載料・制作料により収益をあげている．すなわち，求人広告を増やすとより大きな収益をあげることができる．求人企業側は，活発に応募する会員が多い転職サイトに対して，求人広告を掲載したいと考えるため，転職サイトは応募者を増加させることが最重要である．しかし，現状では転職サイトに会員登録をするものの一度も応募せずに離脱する会員が多い．よって，転職サイトを運営する企業は，離脱傾向がある会員の特徴を探り，離脱しそうな会員に対して適切な施策を講じることが必要である．

離脱予測は，EC サイトやクレジットカードの分野で広く行われており，様々な機械学習手法が用いられている．しかし，それらの研究ではサポートベクターマシンやロジスティック回帰など分類の過程がブラックボックスの機械学習手法が多く使われている．そのため，離脱する会員の特徴を把握することができず，具体的な施策を講じることが難しい．

一方，決定木などの分類が行われる過程が明瞭な機械学習手法を使って，離脱予測を行っている研究もある．これらの研究では，行動ログや会員の登録情報のデータが使われている．しかし，会員には登録された情報以外に経験や資格など様々な特徴を持つはずである．従来の研究では，登録情報からは分からない会員の内在する特徴を見つけることができていない．会員の内在する特徴をつかむことは，適切な施策を講じる上で重要である．

連絡先: 上門 雄也，東京理科大学理工学部経営工学科，7413012@ed.tus.ac.jp

1.2 目的

本研究では職務経歴書を用いて離脱する会員の特徴を抽出し，抽出した特徴と登録情報から離脱予測を行うことを目的とする．職務経歴書とは，会員が自由に記述する文書であり，職務経験や所有資格が記述されている．自由記述文書の職務経歴書を用いることで登録された情報より細かい情報が分かるので会員の内在する特徴までつかむことができる．

2. 関連研究

離脱予測の研究は広く行われているが，問題点がある．本章では，従来の研究と問題点を説明し本研究の特徴を示す．

2.1 Guangli らの研究

Guangli ら [3] は，ロジスティック回帰と決定木を用いてクレジットカードの離脱予測を行った．登録情報と顧客の行動ログを特徴量としており，結果としてどの特徴量が重要か述べている．ロジスティック回帰では，どの特徴量がどのくらい離脱の有無に寄与したかが把握できる．しかし，分類過程がブラックボックスのため離脱の有無を決定づける境界線が分からない．したがって，施策を練ることが難しいとされる．対して，決定木ではどの特徴量がどのように離脱の有無を決定付けたか if-then ルールによって明記されている．これにより，重要な特徴量を理解することができる．しかし，Guangli らの研究の特徴量は会員の登録情報と行動ログだけであるため，会員の内在する特徴を見つけることができない．従来の研究では，会員の内在する特徴が分からないため，具体的な施策を講じることが難しいという問題点がある．

2.2 本研究の位置付け

従来の研究では，予測ができても離脱の有無を決定づける要因が分からなかった．また，その要因が分かる機械学習手法を用いても，特徴量から会員の細かい性質を把握することができ

ない．そのため，離脱を防ぐ施策を講じることが難しいという問題点がある．

本研究では，会員の内在する特徴を抽出し，離脱の有無を決定づける要因を探る．これは，特徴量にテキストデータの職務経歴書を含めることによって可能にする．職務経歴書をテキストマイニングすることで，新たな特徴を抽出することができる．テキストデータを用いて会員の内在する特徴まで抽出するという点で他の研究と区別される．

3. テキストマイニング

テキストマイニングとは，テキストデータを対象としたデータマイニングである．テキストマイニングでは文書の内容を表現するためにベクトル空間モデルを用いる．この章では，日本語の文書ベクトルを作成するための標準的な手法について説明する．

まず初めの手順は，コーパスの作成である．コーパスとは，研究に用いる言語資料のことである．通常，Web ページや新聞などからクローリングして作成される．また，後の手順のために，集められた文書の一部を置換することがある．

コーパスを作成したあと，文書をわかち書きにする．英語の文では，単語ごとに空白で区切られているが，日本語では区切りが存在しない．そこで，ソフトを用いて文書を単語ごとに区切ったわかち書きに変換する．

次に，stop word の削除を行う．stop word とは，分析において確実に意味がないので，取り除かれるべき単語のことである．単語自体に意味をなさない助詞，助動詞やほとんどの文書で使用されている単語がその例である．stop word を削除することで，入力するデータのノイズを減らし，より高い精度の結果を得ることができる．

最後にベクトル空間モデルを用いて，文書ベクトルを作成する．文書ベクトルは，単語の出現頻度を要素としたベクトルである．すべての文書ベクトルを組み合わせると，文書行列によって表される．文書行列の行は文書ベクトル，列はある単語の出現頻度を表している．通常，この文書行列を機械学習の入力データとし，分類などを行う．

4. データセット

本章では，使用するデータについて説明する．本研究では，企業から提供を受けたデータを一部変更して使用した．

4.1 登録情報

会員は，登録時または登録後に履歴書，職務経歴書，自己 PR 書，希望条件を任意で記入する．履歴書と希望条件の記述欄は選択式で，職務経歴書と自己 PR 書は自由記述文書である．履歴書には，年齢や勤務年数など，希望条件には，会員の性格や希望する会社の雰囲気などが項目にある．

本研究では，2009 年 10 月 1 日から 2016 年 8 月 28 日までの履歴書，職務経歴書，希望条件を使用する．自己 PR 書は職務経歴書と内容が類似しており，職務経歴書より情報量が少なかったため，今回は使用しなかった．

また，全てのデータに正解ラベルが振られている．本研究では離脱する会員を「登録して 1 ヶ月以内に一度も応募しない会員」と定義する．よって，1 ヶ月以内に一度も応募しない会員には 1, それ以外の会員には 0 の正解ラベルが割り当てられている．1 ヶ月以内と定義した理由は，登録してすぐに 応募する会員は何度も利用する割合が高く，その一方，1 ヶ月以内に応募しない会員はほとんどが二度と利用しないからである．

4.2 職務経歴書

職務経歴書とは，会員が任意に入力する自由記述文書である．本研究で用いた職務経歴書は 2000 文字以内に限定されている．職務経歴・業務内容・得意分野・アピールポイントなどが記述されており，会員は勤めた企業の数だけ書くことができる．本研究では，職務経歴書を書いている会員から特徴を抽出した．文になっていなかったり，編集中などの不完全なものは分析対象から外した．

5. 提案手法

この章では，提案手法を手順ごとに述べる．本研究では，離脱要因の抽出と離脱予測の 2 つの手法に分かれる．

5.1 離脱要因の抽出

5.1.1 コーパスの作成

はじめに，研究に使うためのコーパスを作成する．本研究では，企業から提供を受けた職務経歴書を用いる．職務経歴書は平文で，2000 文字以内で述べられている．会員の職務経歴書を抽出し，本研究のコーパスとした．また，カナ，数字，アルファベットには半角と全角が混ざっていた．したがって，ライブラリの mojimoji を利用して，カナは全角に数字とアルファベットは半角に変換した．

5.1.2 形態素解析

平文では，機械学習にかけることができないので，形態素解析を行い単語ごとに分割する．本研究では，日本語形態素解析ソフトの Mecab を使用し，文書をわかち書きにした．また，助詞，接続詞などの意味のなさない単語や，形容詞，形容動詞といった会員の特徴がわからない単語は不要のため，名詞のみを抽出した．

5.1.3 stop word

必要のない単語を取り除くため, stop word を作成する．「職務」,「経歴」,「業務」,「内容」といった単語を多くの会員が使用していた．これらは，離脱の有無に関係がないと考えられるため stop word とし，分析対象から外した．

5.1.4 文書行列変換

わかち書きにした文書をベクトルに変換する．文書に出現する単語数と同次元のベクトルを作成する．本手法では，要素を単語の出現頻度ではなく，出現有無のみで表す．すべての次元に単語が割り振られており，文書にその単語が出現していれば 1，出現していなければ 0 とする．すべての文書をベクトル化し，組み合わせることで文書行列を作成する．

5.1.5 出現頻度による特徴量削減

出現頻度が高い又は低い単語は分析に不要である．なぜなら，半数以上に使われるような単語は，正事例と負事例を分ける単語ではなく一般的な単語であるといえるからである．また，逆に出現頻度が低い単語は，使っている会員が少ないためうまく分けられたとしても影響が小さいので施策につながらないと考えられるためである．そこで，本研究では全体の 60%以上の会員が使用している単語と 100 人以下しか使っていない単語を削除した．

5.1.6 重要度による特徴量削減

ランダムフォレストを使って重要度を算出し，重要度が低い単語を削除する．まず，はじめに grid search でランダムフォレストの最適なパラメータを決定する．
求めた最適なパラメータで，ランダムフォレストを 100 回行い，それぞれの単語の重要度の平均値を算出する．grid search では乱数のシード値を 0 で固定したが，ランダムフォレスト

を 100 回行う際は 1 回ずつ 0 から 99 に変更した．本来ならばシード値を変えるごとに grid search をすることが望ましいが非常に時間がかかるため，シード値が 0 のときの最適なパラメータをすべてに適用した．そして，求めた重要度の上位 90%を占める単語以外を削除した．

5.1.7 決定木

削減した特徴量で決定木分析を行う．ランダムフォレストと同様に，grid search で最適なパラメータを決定する．本研究では，求めた最適なパラメータで決定木を行い，離脱する会員と応募する会員の違いのルールを出力する．

5.2 離脱予測

この節では，職務経歴書と登録情報を用いて離脱予測を行う手法を説明する．

5.2.1 前処理

分析を行う前に，登録情報と職務経歴書のデータに対し前処理を行う必要がある．

登録情報のデータには，年齢や転職回数などの数値データと住所などのカテゴリデータがある．分析を行う際，カテゴリデータは数値データに変換する必要がある．one-hot encoding を行い，カテゴリデータを 1,0 の数値に変換する．

職務経歴書のデータは 5.1 で抽出した重要度が高い単語のみ用いる．5.1.4 と同様に，単語数と同次元のベクトルを作成する．次元にそれぞれの単語を割り振り，出現していれば 1, 出現していなければ 0 とする．

5.2.2 職務経歴書との統合

登録情報と職務経歴書のデータを統合する．それぞれのデータには会員 id が含まれているため，会員 id を参照し統合を行う．会員 id が登録情報と職務経歴書の両方に存在している場合，それぞれのデータを統合する．会員 id が職務経歴書には存在せず，登録情報にのみ存在している場合職務経歴書のデータはすべての単語が現れていないということなのですべて 0 にして統合する．また，会員 id が登録情報に存在せず，職務経歴書にのみある場合，登録情報がわからないため分析対象から外す．

5.2.3 重要度による特徴量削減

5.1.6 と同様の操作を行い，ランダムフォレストにより重要度が低い単語を削除する．

5.2.4 決定木

5.1.7 と同様の操作を行い，決定木により離脱する会員と応募する会員の違いのルールを出力する．

6. 結果と考察

6.1 離脱要因の抽出

grid search で求めた最適なパラメータで決定木を行った結果は表 1 に示すとおりである．

表 1: 決定木の結果（職務経歴書）

accuracy	0.66
precision	0.62
recall	0.66

決定木の精度をみると，0.66 と高いとは言えない結果となった．よって，職務経歴書からは離脱を予測することは難しいと考えられる．しかし，本研究の目的は離脱を正確に予測するのではなく離脱の傾向がある特徴が分かれば良い．実際に，木の分類結果をみると離脱傾向がある会員の特徴が分かる．

図 1 の木はノードに書かれている単語が文書に含まれていれば右へ，含まれていない場合は左へ進む．value の値は [負事例，正事例] となっている．すなわち，value の左は応募した会員数，右は離脱した会員数である．根ノードのとき，離脱した会員の数は応募した会員の約 2 倍である．

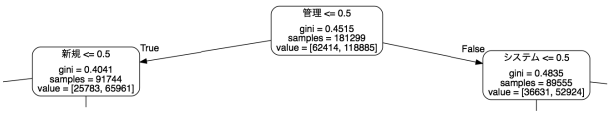

図 1: 決定木（職務経歴書）の根ノード

「管理」が含まれていないとき，すなわち，左のノードへ進んだとき離脱した会員数は応募した会員数に比べ 2.5 倍にまで増える．よって，「管理」という単語が含まれていない会員は離脱しやすい傾向がある．

また，図 2，3 に示すように，ノードに現れた単語をみると「営業」，「開発」，「企画」，「販売」といった内在する特徴がある．これらの単語が現れていると応募している割合が増えていることが分かる．これは，この転職サイトにおいて営業職や開発職の求人が優れているからと考えられる．言い換えれば，それら以外の職種は離脱する傾向がある．なので，営業職や開発職以外の求人を増やすなどの施策を講じればよいと推測される．

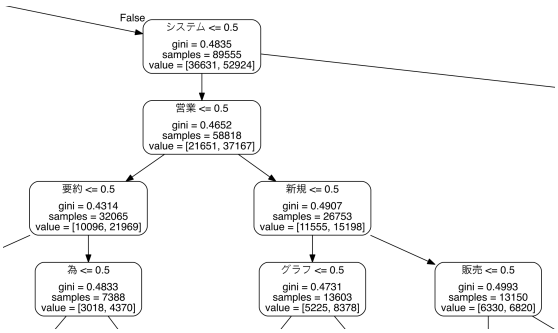

図 2: 内在する特徴「営業」「販売」

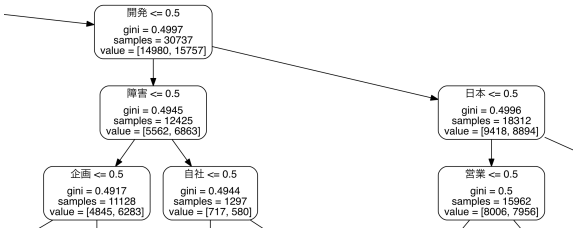

図 3: 内在する特徴「開発」「企画」

次に，離脱傾向が特に強かった箇所を図 4 に示す．

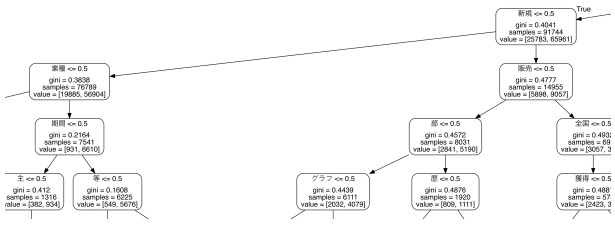

図 4: 離脱傾向が強い決定木の一部（職務経歴書）

図 4 の右上に出現している「新規」は図 1 の左下に出現しているものと同様のノードである．「新規」が出現しておらず，「業種」と「期間」が出現している会員は離脱傾向が強く応募している会員に比べ 10 倍以上も離脱している．これは，離脱する会員は職務経歴書を細かく書いておらず，業種や勤務期間といった大まかな職務経歴しか書いていないからであると考えられる．したがって，会員が細かく書けるようにテンプレートや職務経歴書例を充実させると良いと推測される．

6.2 離脱予測

grid search で求めた最適なパラメータで決定木を行った結果は表 2 に示すとおりである．

表 2: 決定木の結果（統合データ）

accuracy	0.82
precision	0.81
recall	0.82

職務経歴書のみでは，離脱を予測することが困難だったが登録情報を統合することで高い精度で離脱する会員を予測することができた．出力した木を図 5 に示す．

最初の段階では，離脱する会員の数が応募する会員に比べ 3.2 倍であった．根ノードの中身は履歴書の有無であり，履歴書がある会員は右に進み，書いていない会員は左に進む．履歴書を書いていない会員をみると，離脱する会員が 応募する会員の 11.8 倍にまで増えている．したがって，履歴書を書かない会員は離脱しやすいといえる．これは，転職意欲が低い会員は登録のみして履歴書を書くことを後回しにするからと考えられる．

7. 結論

本研究は，離脱を防ぐ施策を講じるために職務経歴書を用いて離脱要因の抽出と離脱予測を行った．手法として，職務経歴書をテキストマイニングすることで会員の特徴を表す単語を抽出した．また，ランダムフォレストで特徴量を削減し，決定木で離脱の要因となる単語を示した．さらに，登録情報と統合することで離脱予測を行った．結果として，職務経歴書だけでは離脱を予測することは困難だったが，「営業」，「開発」などの単語を含む会員は離脱しにくい傾向があることがわかった．これは，営業職や開発職がこの転職サイトでは優れているからと考えられる．よって，その他の職種の求人を充実させることで離脱する会員を減らせるのではないかと推測できる．また，「業種」や「期間」といった単語がある会員は離脱しやすいという傾向があるとわかった．これは，離脱する傾向がある会員は職務経歴書を大まかにしか書いていないためと考えられる．よって，会員が書きやすいように例文を用意するなどの施策を講じればよいと思われる．

さらに，登録情報と統合することで高い精度で離脱が予測できることを示した．

今後の展望として，同じ意味の単語をまとめることでより高精度の予測ができると考えられる．例えば，「店」と「店舗」は同じ意味であるが今回の分析では別の単語として捉えられてしまう．潜在的意味解析により類似した単語をまとめることで，会員の特徴を示す単語が明白になり職務経歴書で予測ができると考えられる．

参考文献

[1] 古川 靖洋. “終身雇用制の現状と人的資源管理”, 産研論集, Vol.38, 94-95, 2011

[2] 中村 天江. “転職経路の「すみわけ」に関する分析—ホワイトカラー正社員の転職活動の実態—”, Works review, vol.6, 166-169, 2011

[3] Guangli nie, Wei Rowe, Lingling Zhang, Yingjie Tian, Yong Shi. “Credid card churn forecasting by logistic regression and decition tree”, Expert Systems with Applications, Vol.38, pp15273-15285, 2011

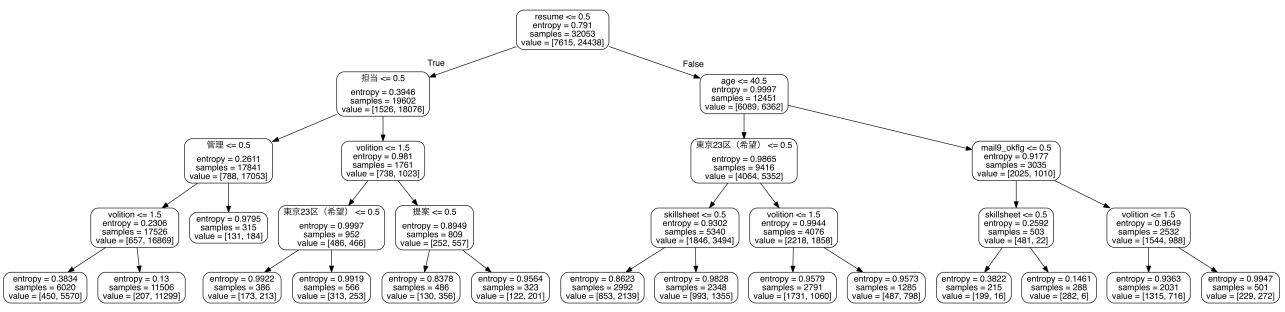

図 5: 決定木（統合データ）