3F3-4

大学中退予防のための相談促進を目的とした機械学習による悩み事アウェアネスの提案

Proposal of awareness support system for one's troubles by machine learning for the purpose of consultation promotion, to prevent a college dropout

川井雄太[*1]
Yuta Kawai

三好康夫[*1]
Yasuo Miyoshi

[*1] 高知大学理学部応用理学科情報科学コース
Course of Information Science, Department of Applied Science, Faculty of Science, Kochi University

Recently, the issue of a college dropout attracts attention, also a management method of improving student's satisfaction and preventing a dropout gives effective results. In addition, it is reported that students who have sufficient communication with teachers are highly satisfied with school. This study is intended to develop the consultation promotion support system between students and teachers, to prevent a college dropout. Specifically, the system classifies mutters on Twitter by machine learning, and performs awareness support to remind a student of one's troubles.

1. はじめに

近年大学中退の問題が注目を集めている. 中退後の学生は非正規就業者, もしくは無就業者が 80%を超え[労働政策研究 2012], 休学者, 留年者同様自殺のリスクが高い[内田 2010]. 大学側も学費が得られなくなることにより経済的損失を被り, 中退率が高いと学位授与・評価機構による評価が下がる. 一方で, 学生満足度を上げて中退率を下げる経営が行われ, 成果を上げている. また先行研究により, 教員とのコミュニケーションが学習意欲を高め, 学生の満足度に影響を与えると報告されている. そこで本研究では, 学生と教員の相談コミュニケーションを促進することで学生満足度を高め, 中退を予防するシステムの開発を目的とする. 本論文においては, 1)悩み事アウェアネス支援機能 2)匿名相談支援機能からなる中退予防システムの提案と, ソーシャルメディアから相談対象となる悩み事を抱えた学生を検出する手法の提案と実験について述べる.

2. アプローチ方法

2.1 エンロールメント・マネジメント

エンロールメント・マネジメントとは, アメリカを中心に行われている中退予防に主眼を置いた大学マネジメント方法である. この手法は, 徹底したデータ分析に裏打ちされた科学的アプローチにより学生の満足度を上げ, 中退を予防するところに特徴がある. 日本では山形大学, 茨城大学等で行われている.

2.2 学生満足度に関する要因

見舘らによると, 大学生活の満足度と学習意欲を押し上げる要因を調査し分析を行った結果, 教員とのコミュニケーション(学習面・生活面の相談回数)が学習意欲を高め, 大学生活の満足度にも影響を与えていた[2008 見舘]. また, 星野らの調査では, 積極的動機がない学生ほど教員とのコミュニケーションにより, 授業満足度や自己努力が改善するなど, 学習意欲の低い学生ほど効果が高い可能性が示唆されている[星野 2006]. そこで本研究では, 学生・教員間の相談コミュニケーションを促進し, 学生満足度を向上させ中退率を下げる事を目的とする.

2.3 相談促進方法

市瀬らの調査[市瀬 2014]によると, 学生が相談しない理由には「悩みは自分で解決できるのが望ましい」「相談機関の敷居が高い」「相談することでイメージが悪くなる」「悩みを話すことへのマイナスイメージ」「相談を周囲に見られたくない」が多かった. また同様に, 求められる社会資源としては「気負わずに小さな悩みで相談できる相談機関」「ネットで繋がれる機能や場」が挙げられていた. このことから, 困っていても学生が相談しないという現実があり, 内野による 1)自身の問題への自覚 2)被援助志向性の低い学生が相談しやすくする工夫, が相談促進に必要である [内野 2010]との指摘は正しく, 上記をサポートすれば相談件数の増加が期待できる. 本研究では内野の指摘に基づき 1)問題への自覚 2)被援助志向性の低い学生の被援助行動を支援するシステムの開発を目的とする.

3. 提案システム

被援助志向性の低い学生を 1)Twitter(http://twitter.com)マイニングによる悩み事アウェアネス支援機能 2)匿名相談支援機能, により相談しやすくサポートする相談窓口アプリの開発を提案する. Twitter を選択したのは, 学生の利用率が高く, また最も本音を発しやすい SNS であるからだ[東京工芸大 2012]加えて生活情報を発信するユーザが多く, 現実の悩みの抽出に適している. 本節では悩み事アウェアネス支援機能と, 匿名相談支援機能について述べる. なお, 利用シーンを考慮し, 学生はスマホアプリとして, 教員は Web アプリとして利用することを想定している.

3.1 悩み事アウェアネス支援機能

図 1 はアウェアネス支援システムを表している. アプリは相談機能のみの使用も可能だが, 基本的にアウェアネス支援機能が作動後に匿名相談支援機能を使用することを想定している.

連絡先:三好康夫, 高知大学理学部応用理学科情報科学コース, 〒780-8520 高知県高知市曙町 2-5-1, 088-844-8346, miyoshi@is.kochi-u.ac.jp

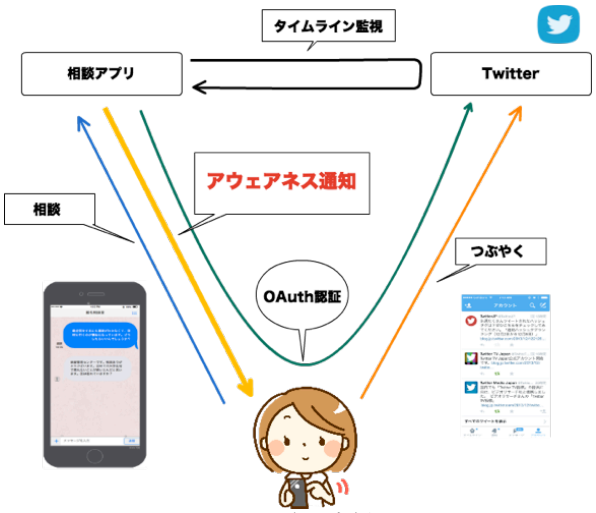

図 1 アウェアネス支援システム

学生はアプリを OAuth 認証により Twitter と連携させる．これは鍵付きアカウントのツイート取得のためである．システムはツイートを加工し，個人情報を削除した上で定期的にサーバーに送信し，解析を行う．その結果，悩み事を抽出したら，クライアント側にプッシュ通知を送信する．これにより，悩み事に気づいていなかった学生，相談すべきか迷っていた学生の相談をサポートできる．Twitter との連携を行っていない学生に対しては，学部，学年，性別などの属性情報をもとに，類似の属性で多い悩みを提示し，悩んでいるのが自分だけでないことに気付かせ，自分が悩んでいる可能性のアウェアネスをサポートし，相談へ繋げる．

3.2 匿名相談支援機能

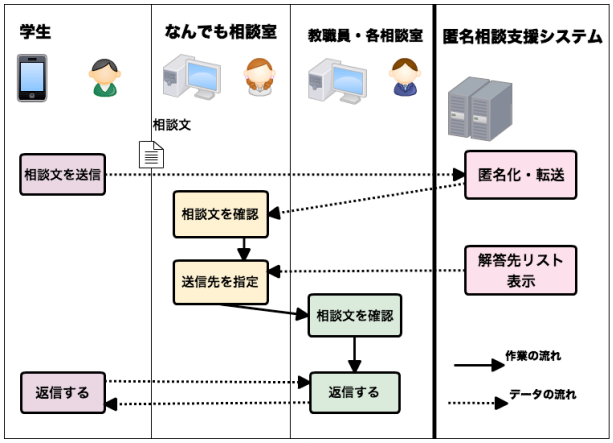

図 2 相談開始後のフロー

図 2 は相談開始後のフローを表している．悩みに気づいたユーザはアプリを通じて相談を行うことができる．被援助志向性の低い学生の相談率を上げるため，相談は匿名で行うことができる．なお，相談はチャット形式で行う．相談はまず一度，何でも相談室など学生の悩みを広く受け付ける部署に送られる．そして相談内容によって保健管理センターや就職支援室など適切な相談機関に振り分けられ，相談者との間で相談が開始される．

ただし，匿名での相談にはメリットとデメリットがある．デメリットは佐藤らによると，相談効果が対面での相談よりも低い[佐藤 2008]．また匿名性を高めるために個人情報を制限して相談した所，情報量の少なさに相談者が解答する上で困難に直面した [仲田 2002]などがある．一方でメリットは，相談自体への抵抗・不安が減少し[佐藤 2008]，信頼できるようになってから心の深層へ話が進むのでなく，いきなり心の深層へ到達しうること[田村 2003]である．このことから，相談のプラットフォームとしては適していないが，相談への敷居が下がり，件数の増加が期待できる．また学内 SNS で相談サービスを行う富山大学では，オンライン上での相談がきっかけでオフラインでの相談件数が約 3 倍に増加しており，オンライン上の相談が手続きとして機能することでオフラインでの相談に繋がる [吉永 2010]．そのため，まずは相談に繋げる，という点が重要であり，それには抵抗と不安が緩和される匿名相談が有効であると言える．

4. 悩み事抽出手法の提案

本研究では悩み事をメンタルヘルスの不調と定義する．メンタルヘルスと就学状況には関係があり，メンタルヘルスに不調があると休・退学・留年しやすい[内田 2010]．また，近年は特に消極的理由での中退が増加している[内田 2009]ことからも，中退者の多くが，在学中に日常的に不調を抱えていたと考えられる．Twitter は本音を発しやすいため，中退者は日常的に不調を呟いており，「授業しんどい」などのツイートが多くなると予想できる．そうしたツイートの中から「大学辞める」など，中退者特有と思われるものを除去した特徴を抽出すれば，メンタルヘルスに不調を抱えた学生を検知できると仮説を立てる．そしてこの仮説のもと，中退者(中退群)の中退前ツイート，中退しておらず健康な在学生(在学群)のツイートの比較により抽出した特徴語を素性とし，機械学習を行う．これにより，中退者を悩み事のある学生として判別できれば，悩み事アウェアネス支援を行うことが可能となる．

本節では訓練データから特徴抽出を行う手法について提案する．なお，学習するのはツイート本文のみとし，投稿日時やリツイート，ネットワークの繋がり等は考慮しない．

4.1 データの前処理

自然言語処理を用いてツイート本文を処理し，特徴抽出に適した形に整形する．ここでは以下のステップでデータ処理を行う．

(1) URL，メンション，ハッシュタグのフィルタリング

今回はツイート本文を用いて実験を行う．そのため，URL やメンション，ハッシュタグといったメタデータは削除し，ツイート本文の意味解析に適した形にする．

(2) 形態素解析

MeCab（ http://taku910.github.io/mecab ）と mecab-ipadic-NEologd（https://github.com/neologd/mecab-ipadic-neologd）を用いて形態素解析を行う．不調が強く現れるのを，名詞，動詞，形容詞，形容動詞とし，その他の品詞は除去する．

(3) 正規化

同じ言葉でも，twitter，Twitter などの表記揺れや，全角半角文字が混在していることがある．そこでアルファベットは全て小文字に変換し，全角アルファベットと全角数字は半角に変換，半角カタカナは全角カタカナに変換して表記を統一する．また活用形も全て原型に変換し，できる限り意味の統一を測る．

(4) ストップワードの定義

「今日」，「する」，「いる」等は多くのツイートに含まれているが，意味解析と分類の精度に負の影響を及ぼす．ここでは國府らにより提案された内容推測に適したストップワード[國府 2013]，と SlothLib（http://www.dl.kuis.kyoto-u.ac.jp/slothlib/）を元にストップワードを定義し，頻出語を除去する．

(5) 単語 N-gram 法

例えば「大学」という単語に注目した時，中退群では「辞める」が共起し，在学群ではしない等予測される．共起関係も考慮すれば精度の向上が期待できるので，単語単位で N-gram 化(1≦N≦7)し，連続する単語を全て一つの単語として切り出す．例えば，「今日 大学 辞める」であれば，「今日」「大学」「辞める」「今日大学」「大学辞める」「今日大学辞める」として切り出す．

4.2 特徴語の抽出

機械学習を行うにあたって，両グループを区別するための特徴が必要になるが，中退群，在学群では共通する属性が多い．そのため，各群の差異を考慮しない Bag of Words は特徴抽出に不向きである．TF-IDF は差異を考慮するが，語数のみを考慮し，日常的に使われる特徴語の抽出には向かない．中退者が日常的に呟く特徴語を抽出するためには，近い属性から特徴を抽出する方法が必要となる．本研究では，そのために，田原・馬の地域特徴語抽出アルゴリズム[田原 2014]を適用する．

(1) 地域特徴語抽出アルゴリズムの適用

田原・馬のアルゴリズムの特徴は対象となる地域から，更に狭い範囲の地域の特徴語を抽出するところにある．例えば，京都市の全 11 区から差異を考慮して京大周辺のユーザの特徴語を抽出する．近い属性情報のグループから差異を考慮して対象のメンバーを抽出するという点で，応用が効くと予想し，今回のケースに適用する．なお，アルゴリズムは以下の通りである．

C_0＝中退群 C_1＝在学群 $|C|=2$ $|D|=100$
$tf(t_i,C_j)=C_j$のユーザが語 t_iを含むツイートを呟いた回数

$$rtf(t_i,C_0)=\frac{tf(t_i,C_0)}{\frac{1}{|C|}\sum_j tf(t_i,C_j)} \quad (1)$$

$$icf(t_i)=\frac{|C|}{cf(t_i)} \quad (2)$$

$$uc(t_i,C_0)=\frac{u(t_i,C_0)}{|U_{C0}|} \quad (3)$$

$$dc(t_i,C_0)=\frac{d(t_i,C_0)}{|D|} \quad (4)$$

$$loc(t_i,C_0)=(1)*(2)*(3)*(4) \quad (5)$$

(1) は語 t_iの C_0における出現頻度を全てのグループでの出現頻度で割った，相対的な出現頻度を表している．C_0での出現頻度が他のグループよりも大きな語は(1)が大きくなる．

(2) は総グループ数を，語 t_iを含むツイートをしたユーザが存在するグループ数で割り，他グループでの語の出現頻度を考慮した出現頻度を表す．その語が出現するグループが少ないほど値が大きくなる．

(3) は C_0のユーザのうち語 t_iを含むツイートを呟いたユーザ数を C_0のユーザの総数で割り，発信ユーザ数を考慮している．これにより C_0の一部のユーザのみが呟く語は値が小さくなる．

(4) は C_0のユーザが語 t_iを含むツイートを呟いた日数を総日数で割り，発信日数を考慮している．これにより突発的な事件やイベントで発信された語は値が下がる．

そして(1)(2)(3)(4)を掛けあわせた(5)の値が高いほど，t_iは C_0において日常的に使われる頻度が高く，特徴度が高い特徴語といえる．

5. 評価実験

実験にあたり，中退群として留学・就職以外で中退した学生 35 名，在学群として重い悩みが見られない在学生 35 名のアカウントを収集した．中退者アカウントは Twitter キーワード検索を用い，退学届の提出が確認でき，中退後に引っ越しが確認できる，友達に別れを告げているなど，実際に退学した可能性が高いアカウントを探した．在学群の学生アカウントはツイプロ(http://twpro.jp)を用いて休・退学・留年しておらず，就学状況に問題がないアカウントを探した．

両グループ計 70 名を訓練データとし，提案手法を用いて抽出した特徴語を素性として交差検証により精度を検証する．なお特徴語の抽出には，中退者は中退前の 100 日間のツイート，在学生は 2015 年 9 月 21 日からの 100 日間のツイートを用いる．規模は合計 76044 ツイートであり，中退群，在学群でツイート数に偏りがないように気をつけた．

5.1 特徴語抽出結果

提案手法を用いて特徴語を抽出した結果を以下の表 1 に示す．

表 1 特徴語 TOP10

順位	中退群	在学群
1	寝る(16)	tweet
2	www(149)	買う
3	食べる(55)	1 日
4	**死ぬ**(136)	twitter
5	学校(315)	終わる
6	帰る(15)	楽しい
7	**無理**(113)	頑張る
8	やばい(20)	好き
9	終わる(5)	なかった
10	バイト(12)	すごい

中退群欄の()内は，その語の在学群での順位を表している．10 位以内に着目すると，中退群では「死ぬ」「無理」と，ネガティブな言葉が 2 つ含まれる一方で，在学群には含まれていなかった．また，ネガティブな言葉の特徴度は在学群では 100 位以下と低い．就学状況とメンタルヘルスの関係を考えると中退群でネガティブな言葉が多いことは不思議でない．そこで 10 位，100 位，1000 位までにネガティブな言葉がどのくらい含まれるか手動でカウントした結果を表 2 に示す．

表 2 上位特徴語に含まれるネガティブ語数

	Top100	Top500	Top1000
中退群	11	46	76
在学群	1	15	22

このように，中退群ではネガティブな言葉が有意($p<0.01$)に多かった．中退群では，少なくとも中退の 100 日前からはネガティブな語が日常的に特徴語として出始めるといえる．そのため，これらを特徴として用いれば中退群と在学群を分類できる可能性が高い．

5.2 ユーザベクトルの作成

上位 1000 の中退群特徴語のうち，ネガティブな言葉に対するユーザベクトルを作成し素性とする．ユーザベクトルはネガティブな言葉に対し，ユーザ u が語 t_i を含むツイートを呟いた日数を，100 日間で実際にツイートのあった日数で割った値として以下のように作成する．

$$w_{ui} = \frac{udc(t_i,u)}{|T_u|}$$

なお，「退学」「退学届」「届」「学校辞める」「学校やめる」「大学辞める」も特徴語に含まれたが，メンタルヘルスの不調のピックアップが目的なため，これらは削除する．以下の表 3 に素性作成に用いた語を示す．

表 3 特徴ベクトル作成に用いる単語

死ぬ	だるい	落ちる	忙しい	クズ	無理無理
ムリ	しんどい	落とす	自殺	ストレス	怒り怒り怒り怒り
痛い	つらい	いたい	負ける	だめ	泣き顔泣き顔泣き顔
やめる	殺す	苦手	無駄	悩む	やめるほしい
怖い	怒り	おかしい	諦める	リスカ	めんどくさい
泣く	つかれる	きつい	耐える	ひどい	迷惑かける
辞める	怒る	死ねる	めんどい	寂しい	落胆の表情
嫌い	疲れる	嫌	邪魔	頭痛い	怒り怒り怒り
朝早い	泣き	辛い	イラ	恐ろしい	行きたくない
ムリ	悩み	休む	いらいら	迷う	寝不足
寝る寝る	地獄	遅刻	腹立つ	死	
こわい	無視	怒り怒り	病む	後悔	

5.3 交差検証

機械学習ライブラリに scikit-learn(http://scikit-learn.org)を用い，Naive Bayes(NB)，Support Vector Machine(SVM)，Random Forest(RF)による精度比較を行った．なお精度は 10-Fold 交差検証を 100 回行った平均値とする．結果を以下の表 4 に示す．

表 4 10-Fold 交差検証によるアルゴリズム比較結果

	NB	SVM	RF
正解率	0.65363	0.69959	**0.72307**
適合率	0.65967	0.77125	**0.77688**
再現率	0.52069	0.54129	**0.63382**
F 値	0.54574	0.60097	**0.66222**

結果は正解率，適合率，再現率，F 値いずれにおいても RF が最も高い精度を示した．

5.4 考察

全ての分類器において，適合率が再現率よりも高い値を示した．ネガティブな言葉のみを素性に用いていることから，中退者の多くが悩みを抱えており，また，この素性を用いることで悩み事を抱えた学生の検出ができる可能性も高い．しかし一方で再現率は低い．これは精神面のファクターのみを用いたために，精神面でネガティブな言葉を発しなかった中退者が検出されなかったと考えられる．飲酒量の増加など行動面のファクターを考慮すると良いかもしれない．また，データ数が少なく，抽出した特徴に偏りがあった可能性，田原アルゴリズムには特定の人物のみが呟く不調語の特徴度を下げる特性があることから，一部の学生が呟く悩みを取り逃がした可能性がある．いずれの場合も，データ数を増やさなくては正確な結果が得られない．今後は今回作成した分類器を用いてより多くの悩みを持つ学生を検索し，より汎用的な素性の設計を行い，再現率を高めることが課題である．

6. おわりに

本論文では機械学習を用いた悩み事アウェアネス支援機能の実装を行った．結果は，ネガティブな言葉を特徴語として機械学習により分類することで，悩みを持つ学生を事前に検知できる可能性が示唆された．より多くのデータから素性を作成することが今後の課題である．

謝辞

本研究の一部は，JSPS 科研費 25330364 の助成を受けた．

参考文献

[労働政策研究 2012] 労働政策研究・研修機構: 大都市の若者の就業行動と意識の展開-『第 3 回若者のワークスタイル調査』から-, 労働政策研究報告書, 2012.

[内田 2010] 内田千代子: 21 年間の調査からみた大学生の自殺の特徴と危険因子, 精神神経学雑誌, Vol.112, No.6, pp.540-560, 2010.

[市瀬 2014] 市瀬晶子, 引土絵未, 李善恵, 大倉高志, 山村りつ, 全海元, 高仙喜, 倉西宏, 尾角光美, 木原活信: 大学生の自殺予防教育プログラムに向けた「悩みとその対処方法」に関する調査-相談することへの抵抗感に着目して-, 人間福祉学研究, Vol.7, No.1, pp.115-127, 2014.

[見舘 2008] 見舘好隆, 永井正洋, 北澤武, 上野淳: 大学生の学習意欲, 大学生活の満足度を規定する要因について, 日本教育工学会論文誌, Vol.32, No.2, pp.189-196, 2008.

[内野 2010] 内野悌司: 大学生の自殺予防, 第 31 回全国メンタルヘルス研究会報告書, pp.22-24, 2010.

[東京工芸大 2012] 東京工芸大: 全国の大学生コミュニケーション調査, https://www.t-kougei.ac.jp/static/file/university-student_communication.pdf, 2012. <2015/10/21 アクセス>

[佐藤 2008] 佐藤広英, 吉田富二男: インターネット上における自己開示ー自己-他者の匿名性の観点からの検討ー, 心理学研究, Vol.78, No.6, pp.559-566, 2008.

[仲田 2002] 仲田洋子: 電子メールを用いた不登校児童支援に関する研究ー不登校児本人とのやりとりを通してー, カウンセリング研究, Vol.35, No.3, pp.276-285, 2002.

[田村 2003] 田村毅: インターネット・セラピーへの招待 心理療法の新しい世界, 新曜社, 2003.

[内田 2009] 内田千代子: 大学における休・退学, 留年学生に関する調査 第 28 報, 茨城大学保健管理センター, 2009.

[吉永 2010] 吉永崇史, 斎藤清二: 富山大学 PSNS を利用したオンライン学生支援, 富山大学総合情報基盤センター広報, Vol.7, pp.14-17, 2010.

[星野 2006] 星野敦子, 牟田光博: 大学の授業における諸要因の相互作用と授業満足度の因果関係, 日本教育工学会論文誌, Vol.29, No.4, pp.463-473, 2006.

[國府 2013] 國府久嗣, 山崎治子, 野坂政司: 内容推測に適したキーワード抽出のための日本語ストップワード, 日本感性工学会論文誌, Vol.12, No.4, pp.511-518, 2013.

[田原 2014] 田原琢士, 馬強: Twitter から有益な日常情報を発見するための特徴語による地域ユーザの検索, 第 6 回データ工学と情報マネジメントに関するフォーラム(DEIM2014)論文集, 2014.