Twitterからの個人の行動に起因するトラブル予測システムの試作

隅田 飛鳥　　服部 元　　小野 智弘
株式会社 KDDI研究所
{as-sumida, gen, ono}@.kddilabs.jp

1　概要

本研究の目的は，個人の行動を要因とするトラブルを事前に警告するシステムの構築である．本稿では，地震や嵐のような災害や，不慮の事故といった多様なトラブルのうち,「高血圧」や「追試」といったトラブルであれば，事前行動から予測可能であろうと仮定し，トラブルを習慣が要因となるものに限定する．このようなトラブルを予測することで，行動履歴に「ラーメンを深夜に食べる」,「運動嫌い」,「ビールを毎晩飲む」などの行動や状態が頻出すれば,「高血圧」というトラブルが起こる可能性が高いことを事前にユーザに告知することが可能となる．

これまでにも本稿と同様に，トラブルを未然に発見するため，トラブルの起きた状況を人手で分析した報告書から要因を抽出し，トラブルの兆しを発見する手法が研究されている [12, 10, 7, 2, 3, 6, 4, 11, 9, 1, 5, 8]. しかしながら，これらの研究では報告書を対象とするため，発生したトラブルの原因や状況を人手で分析する必要があり，個人の行為に起因するトラブルを発見するには不向きである．

本稿では，多様なトラブルに対応し，個人の行動やトラブルが詳細に報告されている文書として，マイクロブログに着目する．マイクロブログは，短い投稿からなるブログの一種であり,「駅前のラーメン屋にきてみた」や「腰痛い」といった，現在の近況や最近の行動が記述されやすい特徴をもつ．そのため，投稿とユーザの行動や状況が一対一に対応しやすく，一般的なブログと比較し，ユーザの行動履歴を得やすい．マイクロブログの中でも，Twitterは，2010年9月14日時点で，世界中で1億7500万ユーザが利用しており[1]，個人の行動に関する豊富な情報をもち，APIを用いたクロールが許可されているメリットがある．本稿では，Twitterからクロールした文書を対象とし，トラブルを予測する．

トラブル予測には，トラブルの要因となる要因候補を発見しなければならない．単純な方法としては，因果関係を用い，トラブルを結果表現として，それに紐づけられている原因表現を，行動履歴から検索する方法が考えられる．しかしながら，トラブル「高血圧」の原因表現「油分や酒の過剰摂取」を行動履歴から検索しても，多くの場合,「ラーメンを食べる」や「ビールを飲む」のように，直接，原因表現が記述されにくく，これらの行動が蓄積された結果，原因表現を示すことが多い．そこで，本稿では，類似した行動の集合をトラブルの要因候補集合とみなし，各候補集合ごとに，トラブルへの起因度を算出する．この起因度を算出するため，具体的には，ユーザ毎にトラブル発生時点をラベリングし，それ以前に発生した要因候補のうち，トラブルが発生したユーザの行動に表れやすい行動に高い重み付けがなされるように機械学習を行う．

[1] http://twitter.com/about

2　提案手法

提案手法を適用する事前段階として，トラブルごとに，ランダムに選択したユーザに対して，トラブルが発生した時点を人手でラベリングする．次に，提案手法では，Step1で，Twitterから行動集合を抽出し，Step2により，ラベリングされる以前の行動集合から，類似した行動の集合を要因候補集合として抽出する．次にStep3により，要因候補集合に対し，トラブルが起きたユーザに特異に発生する行動に高い重み付けがなされるように機械学習により学習を行う．機械学習には，該当のトラブルがユーザに対して発生するか否かを推定するニ値分類問題とみなし，SVMを用いる．

2.1　トラブルラベル付与

まず，提案手法を適用する事前準備として，Twitterデータをクロールし，クロールしたデータの中から，最も新しい1週間のデータを用いて，トラブルが発生した時点を特定するためのラベル付けを行った．Twitterデータは，2010年2月23日〜11月9日 (259日分)の投稿データを対象に，各平仮名をクエリとして検索した結果を，クローリングすることにより収集した．収集したデータ数は401,582,504件で，ユーザ数

表 1: トラブル候補へのアノテーション結果

トラブル	ユーザ数	正例数	負例数
偏頭痛	1609	536	463
風邪	83069	309	691
太った	2496	348	652

は 3,985,160 人であった．収集されたデータのうち，ユーザの最も古い投稿と最も最新の投稿の日時の差をユーザの投稿期間とみなして，平均投稿期間を計算したところ，83 日 19 時間 41 分 1 秒であった．これは収集した期間の約 1/3 に相当し，十分に収集されていないようにみえるが，標本分散値が 80 日 12 時間 56 分 47 秒と高く，ユーザによる差が大きいためだと考えられる．

次に，トラブルを検出するため，事前に用意したトラブル語を含む行動を抽出する．ここでは，トラブル語として，予備実験を元に Twitter 上で投稿されやすい「偏頭痛」,「風邪」,「太った」を対象とする．また，抽出対象には，11/2〜11/9 の間に Twitter に投稿された文書を対象とし，トラブル語を含む文から，ランダムに 1,000 人のユーザを選択し，これらのユーザが投稿したトラブル語を含む文のうち最も古い文を抽出した．抽出した文中には「偏頭痛になりそう」や,「風邪予防しなきゃ」といった実際にはトラブルが発生していないことが明らかな文もあるため，トラブル語の周辺文脈から，トラブルの発生が明らかな場合に正解ラベルを付与し，それ以外の文を不正解とした．ラベル付けを行った結果を表 1 に示す．表の列は左から，トラブル語，該当するトラブル語を含む投稿を 11/2〜11/9 に行ったユーザ数，正例数，負例数を表す．

以降，対象とするデータから，ラベルを付与した投稿時間より前に発生した行動から，ラベル付けされたトラブルが発生するか否かを判定するニ値分類器として学習を行う．なおラベルを付与した投稿時間より前に発生した行動集合から対応するトラブル語を除いた．その結果，0 件になったデータは取り除いた．

2.2 Step1: Twitter からの行動集合抽出

Twitter から行動内容とその発生時間，行動主体者の組を行動とよび，この行動の集合を行動集合として抽出する．Twitter のデータには，ユーザ名，投稿内容，および投稿時刻が紐づけられている．本稿では，ユーザ名を行動主体者，投稿時刻を行動の発生時刻，投稿文書を行動内容と紐付ける．

例えば，ユーザ名「taro」，投稿時間「19:32」，投稿内容「池袋駅周辺でラーメンを食べてきた」であれば，行動主体者は「taro」，行動の発生時刻は「19:32」，行動内容は「池袋駅周辺でラーメンを食べてきた」となる．

2.3 Step2: 要因候補抽出

ここでは，抽出した行動集合から類似した表現で，かつ発生間隔が短い行為の集合を要因候補として抽出する．ここで扱うトラブルは，前述したように，習慣が要因であるものに限定している．習慣を表す表現は，文間の含意関係が成立しているか，あるいは共通の上位概念となる文をもつ含意関係とみなせる．文間含意関係認識では，動詞の含意関係，名詞の上位下位関係などの知識が有益なリソースであることが知られている．本稿では，これらの辞書を用いて，要因候補になりうる表現の組を抽出し，これらの表現の組に含まれる行動集合中の行動の組を習慣とみなす．

まず，動詞間の含意関係辞書の中から含意あるいは同義関係，兄弟語関係となる動詞の組を抽出する．

- A → B の含意関係が成立し，かつ B → A という逆の含意関係の両方が成立する場合，同義関係とみなし，含意関係辞書からこれらの二つの含意関係を除き，同義関係辞書に登録する．
- 発見した同義関係を用い，一語でも同じ語があれば，残りの含意関係を拡張
- 含意関係の中から，共通の語に含意される関係を抽出し，これらの語組に固有の ID を付与する．

次に，名詞の上位下位関係辞書から，上位下位関係，兄弟語関係となる名詞の組を抽出する．名詞には，事態性名詞のように行動を表す語や，食べ物のようにトラブルに起因しやすい語が存在するため，名詞についても要因候補とすることで，動詞だけでは捉えられない情報を補う効果が期待できる．そこで，上位下位関係辞書を用いて，上位下位関係か，あるいは，共通の上位語をもつ関係をであれば，類似した名詞の組だと考え，これらの語組に固有の ID を付与する．

Step1 により得られた行動集合から，名詞あるいは動詞が，辞書から得られた語組に一致すれば，その語組に紐づけられた語組に紐づけられた ID を行動集合に付与する．ただし動詞については，肯否極性が同一であるものを同一の要因候補とみなし，名詞については，格助詞が同一であるものを同一の要因候補とみなす．

2.4 Step3: 要因候補のトラブルへの起因度計算

ここまでで得られた要因候補に 2 でラベル付けした個別のトラブルへの起因度を計算する．具体的には以下のように，トラブルが起こる直前の行動集合から

ユーザ毎に素性ベクトルを生成し，SVM によりトラブルが発生する際の傾向を重みとして計算することで，要因候補のトラブルへの起因度を計算する，

本研究では素性として，ユーザごとに収集した要因候補がある条件 (特徴) を満たすかどうかを一つの素性として表現し，素性ごとに設定された条件を入力の要因候補が満たせば, 対応する素性ベクトルの次元の値に 1 とし, 満たさなければ 0 とする.

INTERVAL トラブルが起こるユーザに特徴的に頻出するイベントであれば，トラブルであろうと仮定し，要因候補毎にイベントが起こる出現間隔の平均を計算する．この平均値の日数と要因候補の ID の組合せに素性を割りあて，相当する素性ベクトルを発火させる

OCC トラブルが発生する人では深夜におき，トラブルが起きないユーザでは，昼間に発生するイベントである場合，そのイベントは，トラブルのトリガーとなっている可能性がある．ここでは，そのトリガーを発見するため，要因候補がおこる出現時刻の平均をとり，深夜帯 (2〜5 時)，夜間 (22〜1 時), 夕食時 (19〜21 時)，夕方 (16〜18 時), 昼間 (13〜15 時)，昼食時間 (12 時)，午前 (9〜11 時), 朝 (6〜8 時) のいずれに当てはまるかをチェックし，要因候補の ID と発生しやすい時間帯との組合せごとに素性を割りあて，対応する素性ベクトルの次元の要素に 1 をセットする

FLUCTUATION 要因候補中でのイベント間の出現間隔の増減

- トラブルが発生する直前に要因候補のイベント間隔が徐々に短くなる傾向があれば，その要因候補はトラブルの要因であろうと仮定する．具体的には，要因候補中のイベントと前後のイベントそれぞれとの間隔を計算し，トラブルに近い順にソートして，$\{t_1 \ldots t_n\}$ とする．n は要因候補に含まれるイベントの数として，以下の式を計算する．

$$\sum_{i=1..n} \frac{t_i - t_{i+1}}{|t_i - t_{i+1}|}$$

この式により得られた数値が負であればトラブル発生時刻に近づくに従ってイベント発生間隔が短くなる傾向を示し，正であればその逆を示す．また 0 であれば，定期的に行われるイベントであることを表す．そこで，この数値が負であれば-1, 正であれば+1, 0 であれば 0 を割りあて，この割り当てた数値と要因候補の ID の組合せ毎に素性ベクトルの次元を割りあて，対応する要素を 1 とした．

表 2: トラブル語ごとのラベル付けした投稿より過去の Twitter データ

トラブル語	投稿数	訓練事例数	評価事例数	平均投稿期間
偏頭痛	865,439	791	198	154 日
風邪	522,650	797	196	68 日
太った	303,122	800	200	35 日

3 実験結果

本提案手法を 2.1 節で述べた 2010 年 2 月 23 日〜11 月 9 日 (259 日分) の Twitter への投稿データのうち，トラブル語ごとにラベル付けされた投稿日より以前のデータを対象に本提案手法を適用する．適用対象データのの投稿数を表 2 に示す．表 2 の各列は，トラブル語，投稿数，平均投稿期間，要因候補数を示す．各トラブル語ごとに，ランダムに選択した 800 件を訓練事例とし，残り 200 件を評価事例として用いた．また訓練事例のうちランダムに選んだ 500 件をパラメータ決定のための予備実験に用いた．

本稿では，形態素解析器に JUMAN[2], 構文解析器に KNP[3]を用いた．SVM には，TinySVM を用い，予備実験の結果から，カーネルは線形カーネルとした．Step2 で用いた辞書として，動詞含意関係辞書には ALAGIN から公開されている動詞含意関係辞書 V.1.2.0[4]に収録されている 55521 対を，上位下位関係には，2010 年 11 月の日本語 Wikipedia のカテゴリと階層構造に上位下位関係抽出ツールを適用して得られた 6,015,759 対を利用した．

本提案手法の有効性を検証するため，比較手法として，以下の 2 種類の手法を適用した．

RAND 評価対象のユーザにトラブルが発生するか否かをランダムに決定

BOW 行動集合中に出現する語ごとに ID を付与し，ID に対応する素性ベクトルを発火させる方法で SVM を学習した結果を適用

表 3 に提案手法と比較手法とを評価した結果を示す．また，本提案手法について，SVM の閾値を様々に設定することにより，精度と適合率のトレードオフを求めた結果を図 1 に示す．表 3 の各列は，評価対象とするトラブル語，手法，正解率，精度，適合率，F 値を示す．ここでは以下の評価尺度を用いた．この結果，風邪と太ったについては，提案手法が最も高い性能が得られた一方，偏頭痛については，BOW より精度が低い結果になった．これは，偏頭痛の要因が現時点で

[2] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[3] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

[4] http://alaginrc.nict.go.jp/

表 3: 本提案手法と比較手法の比較

トラブル語	手法	正解率	精度	適合率	F 値
偏頭痛	RAND(10)	51.10%	52.08%	51.67%	51.87
	BOW	65.66%	64.35%	73.27%	68.52
	本提案手法	54.55%	54.62%	64.36%	59.09
風邪	RAND(10)	50.31%	33.33%	47.58%	39.20
	BOW	65.30%	41.67%	7.58%	12.82
	本提案手法	66.84%	55.56%	7.58%	13.33
太った	RAND	47.50%	31.01%	44.41%	36.52
	BOW	60.50%	33.33%	16.18%	21.78
	本提案手法	64.50%	44.44%	17.65%	25.26

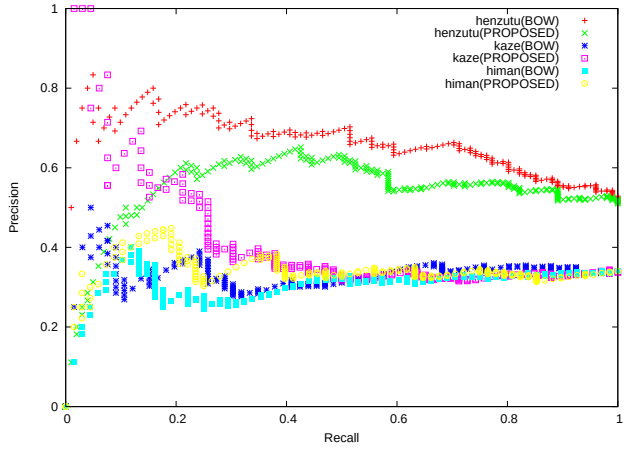

図 1: 再現率と精度のトレードオフ

判明しておらず，行動から予測することがそもそも難しいトラブルであったためだと考えられる．

4 結論

個人の行動を要因とするトラブルを事前に警告するシステムの構築をめざし，単語情報と時間情報との組合せから SVM を用い，自動的にトラブルを予測するシステムを構築した．その結果，約 60%程度の正解率で，適切なトラブルを推定できた．

参考文献

[1] M. A.U Abedin, V. Ng, and L. Khan. Cause identi cation from aviation safety incident reports via weakly supervised semantic lexicon construction. *Journal of Arti cial Intelligence Research*, Vol. 38, pp. 569–631, 2010.

[2] Chidanand Apte, Edna Grossman, Edwin P. D Pednault, Barry K Rosen, Fateh A Tipu, and Brian White. Probabilistic Estimation–Based data mining for discovering insurance risks. *IEEE Intelligent Systems*, Vol. 14, pp. 49–58, November 1999. ACM ID: 630503.

[3] R Bellazzi and B Zupan. Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, Vol. 77, No. 2, pp. 81—-97, 2008.

[4] G. Fung, J. Yu, and W. Lam. News sensitive stock trend prediction. *Advances in Knowledge Discovery and Data Mining*, pp. 481–493, 2002.

[5] R. Girju. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering–Volume 12*, p. 83, 2003.

[6] S. Kogan, D. Levin, B. R Routledge, J. S Sagi, and N. A Smith. Predicting risk from nancial reports with regression. In *Proceeding NAACL '09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 272–280, 2009.

[7] K. Lerman, A. Gilder, M. Dredze, and F. Pereira. Reading the markets: Forecasting public opinion of political candidates by news analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics–Volume 1*, pp. 473–480, 2008.

[8] Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. ARSA: a sentiment–aware model for predicting sales performance using blogs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pp. 607–614, New York, NY, USA, 2007. ACM. ACM ID: 1277845.

[9] Patricia O'Hagan, Edward Hanna, Roy Sterritt, and Paul McKay. Engineering vertical orchestration: From biometric trace events to incident reporting. In *Proceedings of the 14th Annual IEEE International Conference and Workshops on the Engineering of Computer–Based Systems*, pp. 448–458, Washington, DC, USA, 2007. IEEE Computer Society. ACM ID: 1253324.

[10] I. Persing and V. Ng. Semi–supervised cause identi cation from aviation safety reports. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2–Volume 2*, pp. 843–851, 2009.

[11] C. Posse, B. Matzke, C. Anderson, A. Brothers, M. Matzke, and T. Ferryman. Extracting information from narratives: an application to aviation safety reports. In *Aerospace Conference, 2005 IEEE*, pp. 3678—-3690, 2005.

[12] S. Pyysalo, T. Ohta, J. D Kim, and J. Tsujii. Static relations: a piece in the biomedical information extraction puzzle. In *Proceedings of the Workshop on BioNLP*, pp. 1–9, 2009.