2C3-2

Twitterから抽出したユーザの発言に基づく健康状態の推定

Estimation of health condition based on the utterance of the user extracted from Twitter

藤岡 亮太，吉田 稔，松本 和幸，北研二
Ryota Fujioka，Minoru Yoshida，Kazuyuki Mastumoto，Kenji Kita

徳島大学工学部
University of Tokushima Faculty of Engineering

Recently，studies on analysis of medical information using big data is becoming popular．In this study,we propose a method to estimate the current or future health condition of the user based on the remarks of the user from Twitter.The proposed method leads to the prevention of the symptoms of the disease.

1. はじめに

近年，ビッグデータを用いた医療情報の解析に関する研究が盛んになりつつある．例として，医療ビッグデータを利活用し，ビジネスを連携させ，新しいサービスを創出するための研究や新薬や新医療技術等の開発に対する研究などがなされている．

さらに，厚生労働省の報告によると，健康に不安がある人が多く，また，幸福感を判断する際に健康状況が最も重視されることであることがわかっている [1]．よって，健康に関する研究の社会的ニーズは高いと言える．

現在，健康状態を確認する手段は病院での診断以外極めて少なく，仕事などが忙しく，健康診断に行けないような人たちにとって医学知識を必要としない健康状態を確認する手段が必要である，これにより，気軽にマイクロブログを用いて健康状態を確認する手段が提供できる．

本研究では，健康が気になるが，忙しく，健康診断に行けないような人たちが気軽に健康状態がわかるシステムを構築するために，Twitter から現在の健康状態を推定することを目標とし，症状を発言している人物の特徴の有無を調べる手法を提案する．提案手法によって症状の対処や病気の予防につながる．

2. 関連研究

まず，ブログ記事に対する健康アドバイスの自動生成に関して仲村らが提案している手法がある [2]．この研究は，健康情報のうち，食生活と運動に焦点を当て，ブログ記事から食生活と運動に関する情報を抽出し，人間のような言葉でアドバイスを自動生成する技術を開発している．しかし，食生活と運動だけでは健康情報として少ないといった問題点がある．本研究との差異は，本研究では，ブログではなく Twitter を用いるということである．そして，本研究では，健康状態を推定することから，食生活や運動だけでなく，仕事や趣味などさまざまな行動にアドバイスが可能である．

次に，インフルエンザ流行検出のための事実性解析について北川らが提案している手法がある [3]．この研究は，Twitter を用いて，インフルエンザ流行検出のために，話し手の判断や感じ方を表す言語表現であるモダリティを利用した手法を提案している．本研究との差異は，本研究では，病気ではなく健康状態を対象としていることである．また，頭痛，肩こりなどの複数の種類が対象である．

3. 提案手法

3.1 概要

本研究では，Twitter から抽出したツイートデータを TF-IDF で重み付けを行い，cos 類似度を求め，文書同士の類似度から健康状態を推定することを目標とし，症状を発言している人物の特徴の有無を調べる手法を提案する．TF-IDF を用いて症状を発言している人物のツイートデータを特徴付け，複数人でひとつにまとめる．それに症状の発言がある人物やない人物との類似性を一人ずつ調べていくことで，特徴が類似しているかどうかで健康状態が推定可能かどうかを調べる．

3.2 データの取得方法

データの取得方法として，TwitterAPI[4] を用い，症状を発言している人物のタイムライン 500 件 30 人分とそうでない人物のタイムライン 500 件 30 人分をノイズを除去しながら取得する.(図 1)

ノイズもなく，データとして利用可能であると判断出来た場合，テスト用データ，学習用データ，重み付け用データに分ける．症状を検索ワードとして取得したデータは学習用に 20 人，テスト用に 10 人に分け，症状の発言のないデータは重み付け用に 20 人，テスト用に 10 人に分ける．

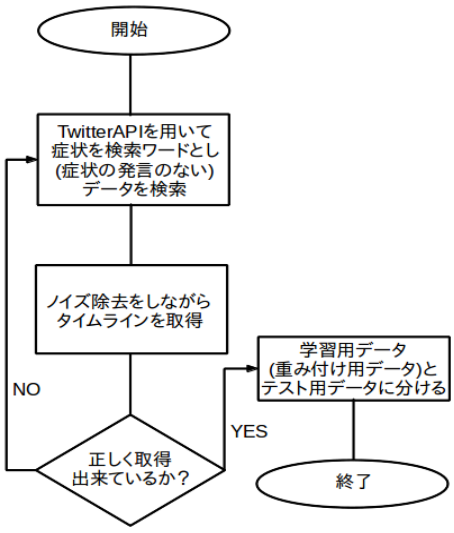

図 1: データの取得方法

3.2.1 TwitterAPI

TwitterAPI とは，Twitter 社が提供しているサービスで、Web 上などから Twitter の機能を利用することができ，これにより不特定多数のツイートを検索，取得することができる．本研究では，TwitterAPI のバージョン 1.1 を使用する．

3.2.2 ノイズの定義

ここでのノイズとは，リンク付き，リツイート，リプライが多くてデータとして不向きなアカウントである．具体的に，タイムライン 500 件の内，リンク付き 50 件，リツイート 40 件，リプライ 150 件よりも多ければノイズとした．

3.3 語の重み付け方法

まず，先ほど取得したデータそれぞれに形態素解析をおこなう．次に，TF-IDF[5] を用いて学習用データ，テスト用データに重み付けを実行する．重み付けをおこなう語を文書の上から順に選択していき，TF を計算する．(図 2)

その後，症状の発言ありの学習用データ 20 人分を 1 つの文書としたものと，症状の発言ありと発言なしそれぞれのテスト用データ 1 人分と，症状発言なしの重み付け用データ 20 人分とで IDF の計算をおこなう．(図 3)

最後に，学習用データとテスト用データそれぞれの TF と IDF の積を計算し，TF-IDF を求める．

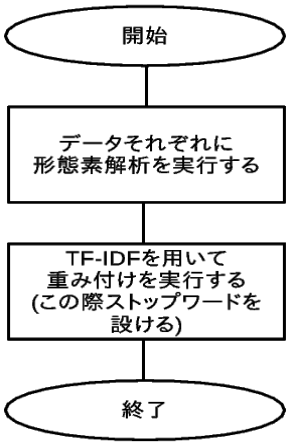

図 2: 重み付け手順

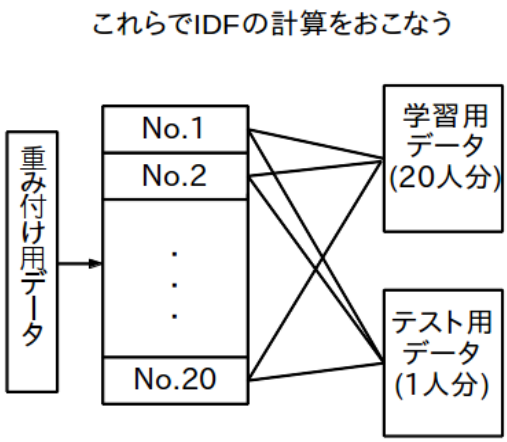

図 3: 重み付け方法

3.3.1 TF-IDF

TF-IDF とは，文書内のある単語がどれくらい特徴があるかを示す指標である．TF と IDF の積を求めることで語の重みの指標となる．すなわち，文書内で多く出現する単語ほど重要で，いくつもの文書で横断的に出現する単語はあまり重要でない．計算式は以下のとおりである．

$$w(t,d) = \frac{tf(t,d)}{S(d)} \quad (1)$$

$$idf(t) = log\left(\frac{N}{df(t)}\right) + 1 \quad (2)$$

$$score(t) = w(t,d) * idf(t) \quad (3)$$

w(t,d)　文書 (d) 内のある単語 (t) の単語の出現割合
tf(t,d)　ある単語 (t) の文書 (d) 内での出現回数
S(d)　文書 (d) 内すべての単語の出現回数
idf(t)　ある単語 (t) の逆文書頻度
N　全文書数
df(t)　ある単語 (t) が出現する文書数

3.3.2 ストップワード

重み付けする際に不要であると判断した語 (表 1) をストップワードとして除去した．

表 1: ストップワード一覧

ストップワード	対象	(例)
1	症状名	肩こり、頭痛
2	文書全体での出現頻度が半分以上の語	@、今日
3	各文書で出現回数が1回の語	アニメやゲームのキャラ名
4	重み上位の語で明らかに重要でない語	下ネタなど

3.4 類似度の算出方法

先ほど重み付けしたそれぞれのテストデータと学習データとの cos 類似度 [6] を求めることで類似度を算出する．(図 4)

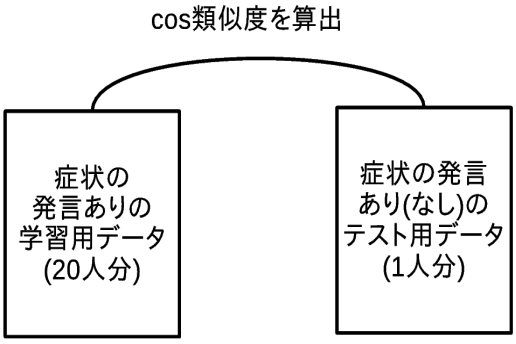

図 4: 類似度の算出方法

3.4.1 cos 類似度

cos 類似度とは，ベクトル同士のなす角度の近さを求めるため，文書同士を比較する場合に使われる手法である，cos 類似度が，1 に近ければ類似しており，0 に近ければ似ていないことになる．

計算式は以下のとおりであり，$\vec{q}$ と $\vec{b}$ はそれぞれ文書内での語の重みベクトルを示す．

$$cos(\vec{q}, \vec{b}) = \frac{\vec{q} \cdot \vec{b}}{||\vec{q}||\,||\vec{b}||} \tag{4}$$

4. 実験と考察

4.1 実験データ

本研究では,「肩こり」,「頭痛」,「吐き気」の 3 つの症状について実験した．使用する実験データは，3 つの症状の学習用データをそれぞれ 20 人分，3 つの症状のテスト用データ (症状の発言あり) をそれぞれ 10 人分，3 つの症状のテスト用データ (症状の発言なし) をそれぞれ 10 人分，重み付け用データ (症状の発言なし) を 20 人分，それらすべてに重み付けをおこなったものである．

4.2 評価方法

実験データを入力とし，cos 類似度を用いて，0〜1 の値を出力する．症状の発言をした人物と発言をしていない人物それぞれのテストデータの cos 類似度を出力し，すべてのテストデータの cos 類似度の平均を求める．

平均を比較し，症状の発言をした人物の平均が高かった場合，症状をもつ人物のツイートには症状固有の特徴があると言える．

これにより，学習データとの類似度から各症状をもっていそうな可能性，これからその病状が現れる可能性を判断できる．つまり，健康状態を推定可能であると評価できる．

5. 実験結果

表 2 に，実験結果を示す．

表 2: 症状ごとの学習データとテストデータの類似度の平均

症状の発言有無 / 症状名	あり	なし
肩こり	0.65	0.60
頭痛	0.69	0.63
吐き気	0.67	0.62

この結果から，総じて症状の発言ありの人物の方がテストデータの平均が高く，症状の発言をしている人物のツイートには固有の特徴があると言える．症状ごとに結果が大きく違うということはなく，3 つの症状が似たような結果となった．しかしながら，症状の発言なしでも発言ありの類似度の平均よりも高い人物も複数存在した．

5.1 考察

実験結果をもとにこのような結果が得られた要因などについて考察する．

1 つ目は，集めた実験データの数が少なかったことが挙げられる．20 人の学習用データに対して，20 人の重み付け用データでは少なく，DF があまり評価されなかったためこういった結果になったと考えられる．

2 つ目は，症状の発言なしの人物に類似度が高い人物と低い人物が混在していたことが挙げられる．以下にテストデータの類似度の一部を示す．

表 3: 「吐き気」の発言なしのテストデータの cos 類似度

アカウント名	Cos類似度
'user1'	0.70
'user2'	0.54
'user3'	0.55
'user4'	0.39
'user5'	0.61
'user6'	0.67
'user7'	0.58
'user8'	0.73
'user9'	0.67
'user10'	0.73

症状の発言ありの
cos類似度の平均=0.67

これはランダムにデータを取得したため起こったので，症状になりそうでない健康な人，もしくは，症状をつぶやかないようなポジティブなツイートが多い人物などを採用するべきであったと考えられる．

5.2 まとめ

本研究では，Twitter から健康状態の推定を行うことを目標に，TwitterAPI を用いて Twitter から症状の発言がある人物とない人物のツイートデータを取得し，形態素解析を行い，TF-IDF による重み付けをおこなった．そして，cos 類似度を計算し，文書同士の類似度を求め，そこから健康状態を推定する実験を行った．

評価実験を通して，全ての症状が，症状の発言ありの人物が症状の発言がない人物よりも類似度の平均が高いという結果となった．これにより，症状を持つ人物のツイートデータには特徴があることがわかった．

一方で，データが少なかったため，TF-IDF を用いた重み付けはあまり有効ではなかったなどの問題もわかった．

5.3 今後の課題

今後の課題として，まず，負例 (症状のない人物) として，本研究では，症状の発言をしていない人物としていた．それを症状になりそうでない健康な人，もしくは，症状をつぶやかないようなポジティブなツイートが多い人物などをツイートを見て採用していくことが挙げられる．

また，データの範囲として，症状の発言した人物のツイートであれば，最新から 500 件のタイムラインを取得していた．それを発言する前後で分けて取得していく，つまり，症状の発言前後でどう類似度が変化するかなどの評価実験を行っていく必要がある．

さらに，実験データの数も増やしていく．以上のことを行い，精度の向上が図れるかを検討したい．

5.4 謝辞

本研究は JSPS 科研費 15K00425,15K00309,15K16077 の助成を受けたものです。

参考文献

[1] 平成 26 年度 8 月分厚生労働省報道発表資料「健康意識に関する調査」，http://www.mhlw.go.jp/stf/houdou/0000052548.html

[2] ブログ記事に対する健康アドバイスの自動生成に向けて，情報処理学会第 77 回全国大会，仲村 哲明，粟村 誉，Yiqi Zhang，荒牧 英治，河原 大輔，黒橋 禎夫

[3] インフルエンザ流行検出のための事実性解析，言語処理学会第 21 回年次大会発表論文集，北川善彬, 小町守, 荒牧英治, 岡崎直観, 石川博

[4] TwitterAPI,http://webnaut.jp/develop/633.html

[5] 自然言語処理，天野真家，石崎俊，宇津呂武仁，成田真澄，福本淳一，オーム社 (2007)

[6] 情報検索アルゴリズム，北研二，津田和彦，獅々堀正幹，共立出版株式会社 (2002)