ソーシャルメディアによる風邪流行の予測

谷田 和章† 荒牧 英治‡ 佐藤 一誠§ 吉田 稔§ 中川 裕志§

† 東京大学大学院 学際情報学府

‡ 東京大学 知の構造化センター

§ 東京大学 情報基盤センター

† punigumi@gmail.com

1 はじめに

ソーシャルメディアと総称される，電子掲示板，ブログ，ソーシャル・ネットワーキング・サービス (SNS) などの Web サービスは，Web サイト制作の知識がなくても簡単に利用することができ，その利用者数を大きく伸ばしてきた．そのなかでも Twitter[1]は，国内だけでもひと月あたり 1000 万人以上の訪問者数がある有名なマイクロブログである．Twitter では，一日あたりの投稿（ツイートと呼ぶ）の数は，国内だけでも 4000 万件以上，世界全体では 3 億件以上になる (2012 年 1 月時点)．これら大量のツイートから実社会を分析することは，様々な目的に活用できると考えられ，注目が集まっている．

風邪の流行は，分析の対象となりうる実社会の事象の一例である．風邪は予防により，その感染を防ぐことができる疾患であるため，流行の早期にその注意を促すことができれば有用だと考えられる．しかし，風邪に対しては，インフルエンザや結核などのような定量的な罹患数の調査は行われていない．ここで，風邪薬の販売量が風邪の流行と強く相関すると仮定することで，この値から風邪流行を推測することができる [1]．しかし，風邪薬販売量は集計されてから公開に至るまでに時間を要するために，従来はこれを風邪流行のリアルタイムな推定に用いることはできなかった．本研究では，リアルタイムに利用できるツイートなどを用いて風邪薬販売量を推定し，風邪流行の把握や予測を行うことを目指す．

2 関連研究

Web 上の情報を用いて感染症の流行を推測する試みとして，これまで様々な方法が提案されている．推測の対象となる感染症としては，その影響の大きさからインフルエンザが取り上げられることが多い．

Ginsberg ら [2] は，検索エンジンの Google[2]に入力される検索クエリのうち，キーワードの検索される頻度を調べることで，その時点の流行の程度が推測できることを示した．キーワードは，それらによる交差検定の結果が悪くなるまで，インフルエンザとの相関が強いクエリを上位から順に選択していくことで得られる．この手法が用いている検索クエリは一般には利用することができないが，Twitter などへの投稿であれば誰でも利用することができ，近年ではそれらを用いた手法が多く提案されている．

Culotta[3] は，先験的に人手によって選択したキーワードを用いて，それを含むツイートの頻度を数えることでインフルエンザの流行推測を行う方法を提案している．ただし，この望ましいキーワードとは別に，望ましくないキーワードも定義しておき，それらが含まれるツイートについては無視することで，実際には罹患と関係しないツイートによる誤った推測を防ぐことができるとしている．また，彼らは，過去の正解データが十分に存在しないにもかかわらず，それらとの相関係数によって推測の性能を評価するのは不十分だとし，誤ったデータを意図的に用いて評価を行う方法についても述べている．

荒牧ら [4] は，ツイートから予め人手によって選択しておいたキーワードを含むものをすべて抽出し，それらを分類器によって実際にインフルエンザの流行と関係するか判別した上で推測を行うことを提案している．

Lampos ら [5] は，二乗誤差に L1 正則化項を加えた損失関数を最小化する Lasso によって推測に用いるキーワードを選択する方法を提案している．この方法により得られるキーワードには重みも与えられるので，流行推測はそれら重み付けされたキーワードのツイー

[1] http://twitter.com/

[2] http://google.com/

トされる頻度を用いて行う．

Achrekar ら [6] は，キーワードの出現頻度だけでなく，当局が発表したインフルエンザの統計値を組み合わせて用いた外部入力付自己回帰によって流行を推測する方法を提案している．ただし，彼らが行った実験では，当局による統計値の発表には 2 週間のタイムラグがあるため，それらは全く用いずにツイートのみを用いた場合のほうが良い精度が得られたと述べている．

これらは，少数のキーワードをあらかじめ決めておく手法と，多数のキーワードを選択する手法とに大別することができる．前者では，適切な単語を見逃している可能性があり，後者では推測精度が同程度であるならキーワードの数は少ないほうが望ましい．本稿では，少数でも効果的なキーワードの選択を試みる．

3　データセット

提案手法では，以下の三種類のデータを用いる．

- 風邪薬販売量
- Twitter のツイート
- 気象情報

これらはいずれも日毎の時系列データである．

風邪薬販売量は，提案手法による推定の正解となるデータである．この統計データには，総務省が公開しているものを利用することができる．ただし，そのままでは日毎の値の差が大きいため，本稿では 7 日間の加重移動平均をして用いる．

ツイートは，単語ごとの日毎の出現頻度として利用する．ただし，各単語の出現頻度には，その出現回数を全単語の出現回数で割った相対頻度を用いる．気象情報としては，気温や湿度などの観測値が利用できる．

これら時系列データは，その期間を訓練期間とテスト期間に分け，それぞれ提案手法の訓練および評価に用いる．先に述べたように，風邪薬販売量の統計データは，薬の購買が行われてから公開に至るまでに少なからぬ時間差があるため，従来これを風邪流行の予測に用いるのは困難だった．一方，ツイートや気象情報などは，ほとんどリアルタイムに利用できるため，それらを用いてその時点の風邪薬販売量を推定することが提案手法の目的である．

4　変数選択による予測手法の提案

提案手法では，まず過去のデータをもとに風邪の流行と関連したいくつかの単語を探し出す．次に，それ

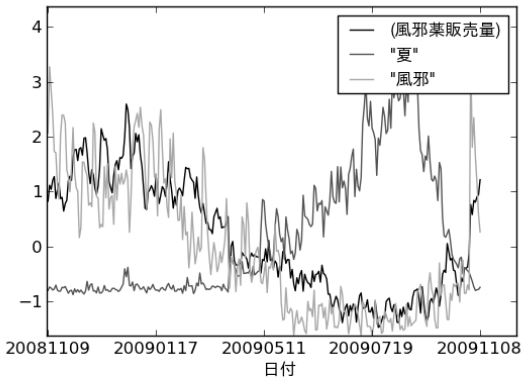

図 1: 利用データの値 (日毎) の例

らの単語がツイートに現れる頻度から，風邪薬販売量を予測する．本節では，説明の都合上，推測について

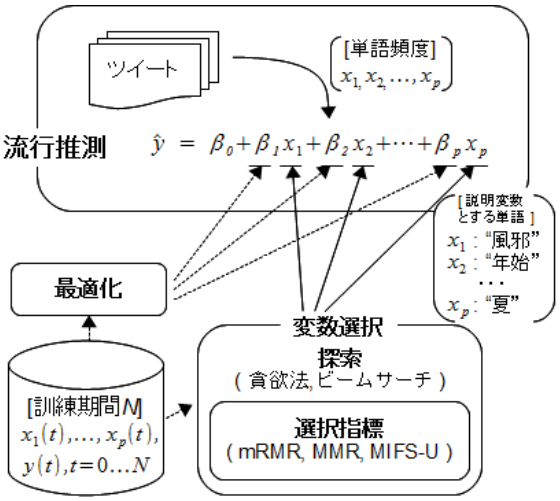

図 2: 提案手法の全体像

先に述べ，次に単語の選択について述べる．

4.1　回帰による推定

提案手法では，風邪薬販売量の推定に線形回帰を用いる．推定する風邪薬販売量は**目的変数**，そのために用いる単語頻度や気象情報は**説明変数**と呼ぶ．時刻 t における説明変数の値として $x_i(t)$ が与えられたとき，時刻 t における目的変数の推定値 $\hat{y}(t)$ は次式で表される．

$$\hat{y}(t) = \ _0 + \ _1 x_1(t) + \ _2 x_2(t) + \cdots + \ _p x_p(t) \quad (1)$$

ここで， $_0$ は定数， $_i$ は各々の説明変数の係数，p は説明変数の個数である．

訓練期間 N のデータが与えられたとき，係数 $_i, i = 1, \ldots, p$ の最適な値は，正解である風邪薬販売量 $y(t)$

と推定値 $\hat{y}(t)$ との目的関数が期間 $N(t = 1, \ldots, N)$ に関して最大となるように定める．二つの時系列データの類似性は，相関係数によって表すことができるため，提案手法ではこれを目的関数として用いる．

4.2 説明変数の選択

ツイートに含まれる単語の種類は非常に多く，それら全ての単語や気象情報を説明変数として用いることは，パラメータの計算時間や，訓練データへの過剰適合などの理由から望ましくない．提案手法では，推定や最適化を行う前に，回帰の説明変数として適切な単語や気象情報を選択しておく．本稿では，説明変数の候補である全ての単語と気象情報とを**素性**と総称する．

4.2.1 選択指標

提案手法では，説明変数とする複数の素性の選択を判断するために，冗長性を考慮した素性選択の基準を応用する．**mRMR**[7] は，用いることができる選択基準のひとつであり，この指標を基にしてある素性 i の良さは次の式で表すことができる．

$$\lambda |r(y, x_i)| \quad (1 \quad \lambda) \frac{1}{|S|} \sum_{x_s \in S} |r(x_s, x_i)| \qquad (2)$$

ここで，r は相関係数，y は風邪薬販売量，x_i は素性 i，S はすでに選択した素性の集合，λ は重みを表す．この式では，ある素性 i について，第一項が正解クラスとの類似性の強さ，第二項がすでに選択した素性との類似性の強さを表す．

また，mRMR と似た考え方に基づく次のような式も選択指標として利用することができる．

$$\lambda |r(y, x_i)| \quad (1 \quad \lambda) \max_{x_s \in S} |r(x_s, x_i)| \qquad (3)$$

$$\lambda |r(y, x_i)| \quad (1 \quad \lambda) \frac{1}{|S|} \sum_{x_s \in S} |r(y, x_s)| \cdot |r(x_s, x_i)| \qquad (4)$$

これらはそれぞれ，**MMR**[8]，**MIFS-U**[9] を基にした指標である．

4.2.2 候補からの探索

選択指標によって評価を行う素性の組合せを決めるためには，探索法を用いる．探索法のひとつである**貪欲法**では，次のようにして素性を選択していく．

1. すべての素性について指標を評価し，その値が最も大きくなる素性を選択する．
2. 1. で選択した素性を選択済み素性集合に加える．
3. 選択済み素性があらかじめ決めた数になるまで 1. から繰り返す．

ビームサーチによる変数選択では，貪欲法と同様に選択指標を用いて単語を順に選んでいく．ただし，探索の各繰り返しにおいて，得られる選択済み素性集合は貪欲法では一通りであったが，ビームサーチではある任意の数 (ビーム幅と称する) の集合を作る．素性の選択は次のようにして行う．

1. 各選択済み素性集合について，新たに 1 つ素性を加えたときの指標の値をすべての組合せについて計算する．
2. 選択済み素性と追加素性のペアのうち，指標の値が上位のものビーム幅数分について新たな選択済み素性集合のリストとして取得する．
3. 各選択済み素性集合内の素性があらかじめ決めた数になるまで 1. から繰り返す．

ビーム幅が 1 の場合のビームサーチによる選択は，貪欲法による選択と同様の結果になる．

5 評価実験

本節では，次の期間のデータセットを用いて提案手法を評価する．

2008 年 11 月 09 日から 2010 年 07 月 04 日
ただし，次の期間はデータがないため除く．
2009 年 03 月 05 日から 2009 年 04 月 18 日，
2009 年 09 月 20 日から 2009 年 11 月 01 日，
2010 年 03 月 26 日から 2010 年 04 月 18 日．

2011 年 06 月 05 日から 2011 年 08 月 31 日
ただし，このうち十数日分のデータが取得できなかったため，それらの日を除く．

時系列データの期間は，変数選択や最適化を行う訓練期間と推定精度の評価を行うテスト期間に分けられる．今回は，訓練期間を 2008 年 11 月 09 日から 2009 年 11 月 08 日までの一年間，テスト期間を残りの期間とする．

表 1 に，選択指標として mRMR を用いたとき，選択された素性による推定精度をビーム幅ごとに示す．推定の精度は，推定値と正解データとの相関係数によって評価する．表中では，選択指標の重み λ を 0.1 から 0.9 までステップ幅を 0.1 として試したとき，訓練期

間の相関係数 R が最も大きくなるものを示してある．また，ビーム幅が 1 より大きく説明変数集合の解が複数存在する場合については，そのうち訓練期間の相関係数 R が最も大きいものを示してある．

表 2 には，提案手法に加えて，人手によって”風邪”という単語を選んだ場合，365 日前の値を推定値とした場合，Ginsberg らの手法を用いた場合の結果を示してある．Ginsberg らの手法は，相関の強い単語を上位から順に足しあわせて推定を行うものである．この手法では，表中の例では上位 3 つの単語が用いられているにもかかわらず，相関が最も強い単語を一つだけ用いた場合と比べて訓練期間の相関はあまり増えていない．一方，提案手法では，同様に少数の単語を用いているにもかかわらず，一つの単語を用いた場合と比べて訓練期間での相関が大きく向上している．また，テスト期間においても提案手法は，その他の手法と比較して高い精度で風邪薬販売量を推定している．

表 1: パラメータ毎の推定精度

		mRMR		
素性数	ビーム幅	λ	訓練 R	テスト R
1	-	-	0.886	0.331
2	1	0.6	0.920	0.453
	20	0.6	0.922	0.709
	40	0.6	0.927	0.717
3	1	0.6	0.931	0.678
	20	0.6	0.943	0.881
	40	0.6	0.944	0.880
4	1	0.6	0.937	0.391
	20	0.6	0.956	**0.894**
	40	0.6	0.956	**0.894**
5	1	0.6	0.943	0.223
	20	0.6	0.962	0.883
	40	0.6	0.962	0.883
6	1	0.5	0.950	-0.205
	20	0.6	0.966	0.854
	40	0.6	0.966	0.853
7	1	0.5	0.954	-0.378
	20	0.6	0.967	0.824
	40	0.6	0.967	0.823

6 おわりに

本稿では，ツイートや気象情報などリアルタイムにインターネットを通じて入手できるデータから風邪薬販売量を推定する方法について述べた．推定に用いるデータは，素性選択基準を探索法によって評価することで選ぶことができる．実験では，少ない種類のデータから高い推定精度が得られることを示した．また，本稿では，提案手法を風邪の流行を推測するために用いたが，例えば市況や世論などの対象にも本手法を利用できる可能性がある．

表 2: 推定精度の比較

方法	素性数	訓練 R	テスト R
人手 (”風邪”)	1	0.832	0.734
自己回帰 (365 日前)	1	0.902	0.885
Ginsberg(2009)	(3)	0.896	0.320
訓練 R 最大	1	0.886	0.331
貪欲法&MIFS-U	4	0.951	0.788
ビームサーチ&mRMR	4	0.960	0.894

参考文献

[1] SF Magruder, Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease, Johns Hopkins APL technical digest Vol.24, 2003.

[2] J Ginsberg, M H Mohebbi, R S Patel, L Brammer, M S Smolinski, L Brilliant. Detecting in uenza epidemics using search engine query data, Nature Vol.457 (19), 2009.

[3] A Culotta, Detecting in uenza outbreaks by analyzing Twitter messages, arXiv:1007.4748v1 [cs.IR], 2010.

[4] E Aramaki, S Maskawa and M Morita, Twitter catches the u: Detecting in uenza epidemics using Twitter, Proceedings of the Conference on Empirical Methods on Natural Language Processing, 2011.

[5] V Lampos and N Cristianini, Tracking the u pandemic by monitoring the social Web, IAPR 2nd Workshop on Cognitive Information Processing (CIP), 2010.

[6] H Achrekar, A Gandhe, R Lazarus, SH Yu and B Liu, Predicting u trends using Twitter data, IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2011.

[7] H Peng, F Long and C Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, Pattern Analysis and Machine Intelligence Vol.27 (18), 2005.

[8] J Carbonell and J Goldstein, The use of MMR, diversity-based reranking for reordering documents and producing summaries, Proceedings of the SIGIR, 1998.

[9] N Kwak and CH Choi, Input feature selection for classi cation problems, Neural Networks Vol.13 (1), 2002.