Twitterによる風邪流行の推測

谷田和章[1*]　荒牧英治[1]　佐藤一誠[1]　吉田稔[1]　中川裕志[1]
Kazuaki Tanida[1]　Eiji Aramaki[1]　Issei Sato[1]　Minoru Yoshida[1]　Hiroshi Nakagawa[1]

[1] 東京大学
[1] The University of Tokyo

Abstract: Many studies pay much attention to information extraction from Twitter, which is a popular microblogging service. Among them, Twitter based infectious disease surveillance is promising, because real time nature of Twitter is suitable for the surveillance. This paper proposes a new method to detect epidemics of cold by counting frequency of some words included in tweets. In our approach, we select these words by criteria based on feature selection method such as mRMR and a searching algorithm such as beam search.

1　はじめに

インターネット上で個人が文章を簡単に公開できるサービスとして，ブログやマイクロブログなどが存在する．そのなかでも，Twitterは，一億人以上のユーザが利用する，人気のあるマイクロブログのひとつである．Twitter上での発言 (以降，ツイートと呼ぶ) の数は一日あたり数億件もあり，これらの大量のツイートを利用して実社会を分析することができれば，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．例えば，風邪は早期の予防により，感染を防ぐことができる疾患であるにもかかわらず，インフルエンザや結核などのように定量的な罹患数の調査が行われていない．そこで，本研究では，ツイートから風邪流行の把握を行うことを目指す．

提案手法の基本的なアイデアは，ツイート中から出現頻度が風邪の流行と相関したいくつかの単語を選択し，それらを用いて風邪の流行を推測することである．ただし，大量の単語から風邪の流行を予測することにはコストが掛かるため，なるべく少ない単語から流行を予測できることが望ましい．本研究では，この問題を素性選択の問題の応用とみなし，minimum Redundancy Maximum Relevance (mRMR) などの基準を用い，ビームサーチによって単語を選択する．実験では，既存研究のように先験的に単語を決めた場合や，正解データとの相関が強い単語を単純に用いた場合に比べ，高い精度で風邪の流行を推測できることを示す．

以下，2節では関連研究について述べ，3節で利用するデータについて述べる．4節では，提案手法について説明し，5節で提案手法を評価する．最後に6節で，本稿のまとめを述べる．

*連絡先：東京大学大学院学際情報学府
東京都文京区本郷7丁目3-1
E-mail: punigumi@gmail.com

2　関連研究

Twitterから情報の抽出を試みた研究には様々なものが存在し，榊らによる地震の検知 [1] や，Bollenらによる株価の値動き予測 [2] などをはじめ，様々な問題がターゲットとされてきた．なかでも，インフルエンザをはじめとした感染症の流行を推測する問題は，Twitterとの親和性が高く，提案手法をはじめ，多くの研究が存在する．

荒牧らは，分類器によってインフルエンザの罹患に関するツイートのみを抽出し，その数から流行を推測する手法を提案している [3]．Achrekarらは，当局が発表したデータとツイートを用いる外部入力付自己回帰によって，翌週の患者数を予測する手法を提案している [4]．これらの手法では，先験的に決めた単語 (”influenza”や”flu”など) の有無によってインフルエンザと関係するツイートを選別しているが，本研究では単語を目標データおよびそれら自身との相関に基づいて選択する．

Ginsbergらは，Googleに入力された検索クエリを用いた方法により，実際のインフルエンザ患者数を高精度で推測できたと報告している [5]．彼らの手法では，推測に用いられる単語として目標データと強く相関するものが選ばれるが，単語同士の相関は考慮されておらず，インフルエンザ流行のある特徴に関して十分に推測できない可能性がある．検索クエリを用いたものでは，他にPolgreenらの研究 [6] などがある．これらの手法は，ユーザのインプットをセンサとして用いるという点では先に述べた研究と共通しているものの，ツイートとは異なり検索クエリは一般に利用可能なデータではない．

3　利用データ

本節では，後の節で用いるデータについて述べる．提案手法では，いくつかの種類のデータを元に，その時点での風邪流行を推測する．このとき，推測する風邪流行を**従属変数**，元となるデータを**説明変数**と呼ぶ．後に述べる変数選択，最適化，および実験などでは，それらのデータはある期間の時系列データとして用いられる．これらの時系列データは，訓練期間とテスト期間に分けることができ，変数選択や最適化など，推測精度の評価以外では訓練期間を使用する．

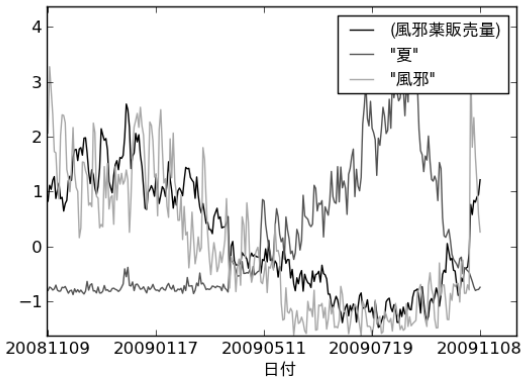

図 1: 利用データの値 (日毎) の例

説明変数の候補となるデータとして，本稿ではツイート中の単語ごとの日毎の出現頻度を用いる．ただし，Twitter の利用者数はこれまで速いペースで増加していることなどから，たとえば今年の 1 日あたりのツイートの数は一昨年の値と大きく異なる．そのため，単語の頻度としては，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．

また，ほかにも最近の値が利用できるようなデータであれば，説明変数の候補となりうる．平均気温などの気象情報は，その一例である．

しかし，従属変数としたい風邪の流行度合いは，調査が行われておらず，これをデータとして用いることができない．そこで，それに代わる指標として，ここでは風邪薬の販売量 (総務省統計局が公開) を用いることにする．風邪薬販売量には，風邪の流行度合いと非常に強い正の相関があると仮定する [7].

風邪薬販売量は，薬の購買が行われてから集計され公開に至るまでに時間差があるため，従来はこれを風邪流行の予測に用いることはできなかった．本稿では，この値をツイートからリアルタイムに推測することで，風邪の流行度合いを示す．

4　風邪流行の推測

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，それらの単語がツイートに現れる頻度から，風邪の流行度合いを推測する．本節では，説明の都合上，推測について先に述べ，次に単語の選択について述べる．

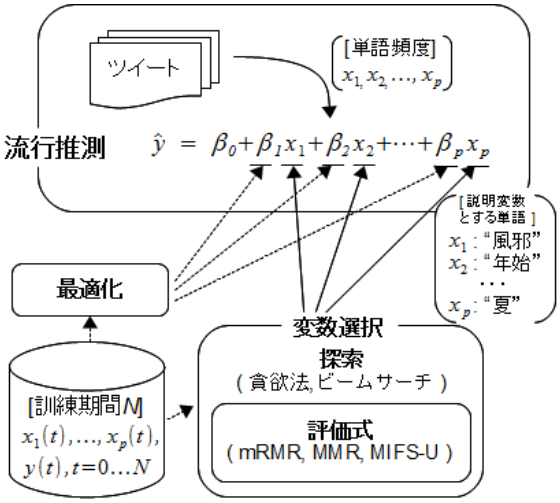

図 2: 提案手法の全体像

4.1　推測

提案手法では，風邪流行の推測には回帰分析の手法を利用する．前に述べたように，推測される風邪流行は**従属変数**，そのために用いられる単語頻度などのデータは**説明変数**と呼ぶ．本稿では，回帰分析として線形回帰を用いる．そのモデルは従属変数 Y と説明変数 $X_i, i = 1, \ldots, p$ について次式で表すことができる．

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \quad (1)$$

ここで，β_0は定数,β_iは各々の独立変数の係数, p は従属変数の個数,ϵ は誤差である．上の式に従うと，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，従属変数の推定値$\hat{y}(t)$ は次式で表される．

$$\hat{y}(t) = \beta_0 + \beta_1 x_1(t) + \beta_2 x_2(t) + \cdots + \beta_p x_p(t) \quad (2)$$

4.1.1　最適化

期間 N において最適な係数β_iは，正解データである風邪薬販売量 $y(t)$ と推定値$\hat{y}(t)$ との間の誤差が最小となるように定める $(t = 0, \ldots, N)$．二つの時系列データの類似度を相関係数 R で表すことにすると，誤差を

最小にする β_i は $R(y;\hat{y})$ を最大にするものである．相関係数 R は次式で与えられる．

$$R(a;b) = \frac{\sum_t (a(t)-\bar{a})(b(t)-\bar{b})}{\sqrt{\sum_t (a(t)-\bar{a})^2}\sqrt{\sum_t (b(t)-\bar{b})^2}} \quad (3)$$

ただし，$\bar{x}$ は $x(t), t = 0, \ldots, N$ の標本平均とする．

提案手法では，実際の推測に先立って訓練期間 N において係数 β_i を定めておき，それを用いてリアルタイムにデータから風邪薬販売量を推測する．

4.2 変数選択

回帰分析では，従属変数の値をいくつかの説明変数から推測するが，その前に説明変数として適切な単語を選択しておく必要がある．全ツイートに含まれる単語の種類は非常に多く，それら全ての単語の頻度をそれぞれ説明変数とすると，パラメータの計算時間の増大や，訓練データへの過剰適合などの原因により，適切な回帰のパラメータ推定が困難となる．

回帰の説明変数が一つの場合には，頻度が風邪薬販売量と最も相関する単語を選択する．このとき選ばれるものは，直感的にも関係性が理解しやすい，例えば「風邪」のような単語であろう．しかし，一つだけでなく複数の説明変数を用いることで，風邪薬販売量の特徴がより良く表わされ，より適切な推測ができる可能性がある．

説明変数として，二つ以上の単語頻度を用いる場合，風邪薬販売量との相関が高い単語を単純に上位から順に選択することは適切ではない．たとえ風邪薬販売量との相関が強い単語が存在したとしても，その単語とすでに選択されている単語との間に強い相関があれば，それらは風邪薬販売量の同じ特徴を表しているといえる．そのため，これらの単語を説明変数としても，推測の精度の向上には結びつくとは限らない．そればかりか，むやみに説明変数を増やすことは，先に挙げたオーバーフィッティングなどの原因ともなる．

4.2.1 選択の評価式

複数の単語の選択は，候補となる単語同士の相関を選択の基準に含めることで実現する．そのために，素性選択の方法である minimum Redundancy Maximum Relevance (**mRMR**)[8] を応用することができる．mRMR では，次のような式の繰返しによって，クラス分類に用いる代表的な素性を選択していく．

$$I(C;f_i) - \frac{1}{|S|}\sum_{f_s \in S} I(f_s;f_i) \quad (4)$$

ここで, I は相互情報量, C はクラス, f は素性, S はすでに選択した素性の集合を表す．この式では，ある素性について，第一項がクラスとの相関の強さ，第二項がすでに選択した素性との相関の強さを表す．繰返しの度にすべての素性について式の評価を行い，値が最も大きくなる素性を新たに選択する．この式を単語選択に利用する場合，次のような式で表すことができる．

$$|R(y;x_i)| - \frac{1}{|S|}\sum_{x_s \in S} |R(x_s;x_i)| \quad (5)$$

ここで, R は相関係数, y は風邪薬販売量, x は単語頻度, S はすでに選択した単語の集合を表す．ここでは，相関の大きさを表す方法として相関係数を用いた．風邪薬販売量と単語頻度との相関は，過去のデータに基づいて算出する．

この mRMR に基づいた評価式は，次のように拡張することができる．

$$\lambda|R(y;x_i)| - (1-\lambda)\frac{1}{|S|}\sum_{x_s \in S} |R(x_s;x_i)| \quad (6)$$

この式では，重み係数 λ によって，目的データとの相関 (第一項) および選択済みの単語の相関 (第二項) のどちらを重要とみなすかを調整することができる．本稿では，この拡張を **mRMR'** と称する．

また，mRMR' と似た考え方に基づく次のような式も単語選択に利用することができる．

$$\lambda|R(y;x_i)| - (1-\lambda)\max_{x_s \in S} |R(x_s;x_i)| \quad (7)$$

$$|R(y;x_i)| - \lambda\sum_{x_s \in S} |R(y;x_s)| \cdot |R(x_s;x_i)| \quad (8)$$

これらはそれぞれ，Maximal Marginal Relevance (**MMR**)[9], Mutual Information Feature Selector under Uniform information distribution (**MIFS-U**)[10] による選択式である．選択済みの単語との相関について，MMR ではそのうち最も強い相関のものだけを考慮し，MIFS-U では目的との相関の強さで重み付けた相関の合計値を利用する．

4.2.2 探索法

これまでに述べた変数選択の評価式を用いることで，目的に適した単語を大量の候補のなかから選択することができる．単純な変数選択式の利用方法である**貪欲法**では，式を繰返し評価して，評価のたびに値を最大にする変数を一つずつ選択していく．アルゴリズム 1 に，貪欲法による単語選択の擬似コードを示す．貪欲法で選ばれる単語は，いずれも適度に独立ではあるが，それらを変数とした回帰における推定の精度は必ずし

も最大ではない．換言すれば，繰返しの各評価において値が最大ではない単語を選択したほうが，選択される他の単語と合わせたときに目的の値を良く近似する場合があり得る．

アルゴリズム 1: 貪欲法

```
            getSelectValue(c, X; s):
                変数選択の評価式(式(5)～(8))
                c: 正解データ
                X: 単語
                s: 選択した単語の集合
<input>     S_all(X_1, X_2, ..., X_K):
                説明変数候補の全単語(K個)
            p:  イテレーション最大回数
            C:  正解データ
                (風邪薬販売量の時系列データ)
<output>    S_sel(X_1, X_2, ..., X_p):
                選択した単語の集合

  copy S_all to S_rem;
  for i = 1 to p {
    X_sel = argmax_{X∈S_rem}(getSelectValue(C, X; S_sel));
    add X_sel to S_sel;
    remove X_sel from S_rem;
  }
```

ビームサーチによる変数選択は，貪欲法と同様に，評価式を用いて単語を順に選択していく．各反復において，それまでに選択された単語の組合せは，貪欲法では 1 通りであったが，ビームサーチでは事前に決めた数 (ビーム幅と称する) だけ組合せを保つ．反復ごとの単語選択では，それまでに選択されているビーム幅 N 個の単語組合せのそれぞれについて新たに組み合わせる候補となる単語の評価を行い，うち上位 N 個の組合せをその反復における選択結果とする．アルゴリズム 2 に，ビームサーチによる単語選択の擬似コードを示す．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．

アルゴリズム 2: ビームサーチ

```
            getSelectValue(c, X; s):
                変数選択の評価式
<input>     S_all(X_1, X_2, ..., X_K):
                説明変数候補の全単語(K個)
            p:  イテレーション最大回数
            b:  ビーム幅
            C:  正解データ
<output>    S_sel(X_1, X_2, ...)[i][j]:
                i = 1, ..., p,  j = 1, ..., b
                選択した単語の組合せのリスト

  getEstValue = function(c, s) {
    if s == φ {
      return 0;
    }
    pop last added X from s;
    return getEstValue(c, s) + getSelectValue(c, X; s);
  }  //組合せの評価

  S_sel[0][1] = φ;
  for i = 1 to p {
    S_cand = φ;
    for j = 1 to b {
      S_rem = {X|X ∈ S_all \ S_sel[i-1][j]};
      for X ∈ S_rem {
        copy S_sel[i-1][j] to s_tmp;
        add X to s_tmp;
        add s_tmp to S_cand;
      }
    }  //評価する単語組合せS_cand取得
    j = 1;
    for s ∈ S_cand order by getEstValue(C, s) DESC {
      S_sel[i][j++] = s;
      if j > b {
        break;
      }
    }  //上位b件の単語組合せ取得
  }  //1～p個の単語組合せを各々b通り取得
```

5 実験

本節では，過去のデータを用いて提案手法を評価する．

5.1 データセット

実験には，次の三種類の時系列データを用いる．

風邪薬販売量 正解データ．7 日間の加重移動平均をして用いる．

単語ごとのツイート中の出現頻度 説明変数の候補．各単語はツイートを形態素解析することで得る．なお，名詞，動詞，形容詞，形容動詞，副詞，連体詞以外の品詞として用いられる単語は省いた．

平均気温などの気象情報 説明変数の候補として用いる．これらはいずれも日毎の値である．今回，これらデータの共通して利用できる期間は，以下の通りである。

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．

5.2 結果

実験では，まずデータを訓練期間とテスト期間に分ける．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．訓練では，訓練期間から回帰の説明変数とする単語もしくは気象情報を選択し，またそれらの係数も決定する．テストでは，テスト期間における風邪薬販売量と回帰による推定値との相関の強さを評価する．

風邪薬販売量と推定値との相関が強いほど推定の精度は高いといえ，その値は相関係数で与えられる．

表 1 に，変数選択の各評価式 (6)(7)(8) を用いた貪欲法による結果を示す．実験では，評価式 mRMR' と MMR では重みを 0.1 から 0.9 まで 0.1 ずつ変え，MIFS-U では重みを 1.0 とした．表には，組合せの単語数が 1 から 7 つの場合において，訓練期間の相関が最も強くなる重みの結果のみを示した．また，比較対象として，単語を「風邪」とした場合 (人手)，Ginsberg らの手法を用いた場合 (Google) の結果も示している．

各評価式について，ビーム幅を 40 としたビームサーチによる結果を表 2 に示す．貪欲法の場合と同じく mRMR' と MMR では 0.1 から 0.9 までの 9 通りの重みについて実験を行い，表にはそのうち訓練期間の相関が最も強くなる結果のみを単語数ごとに示した．

5.3 考察

実験結果のうち，最もテスト期間での相関が大きく，推測の精度が良かったのは，ビームサーチで mRMR' を用いた場合 (λ=0.6, 単語数=4) であった．この単語組合せでは，人手により先験的に単語”風邪”を説明変数とした場合に比べ，相関係数の差が訓練期間では 0.124，テスト期間では 0.160 と大きく上回る．

貪欲法でのテスト期間の相関の多くは低い値となったが，これは最初に選択される単語”ませ”の影響による．”ませ”は訓練期間では最も相関が強いが，テスト期間ではあまり相関がない，ノイズのような単語である．

訓練での相関が高い単語をそのまま上位から順に説明変数とする Ginsberg らの手法では，ここでは 3 つの単語が用いられているにもかかわらず，1 つのみの場合 (”ませ”) と訓練期間の相関がほとんど同じである．一方，提案手法では，少ない単語数であっても相関が高くなる単語を効果的に選択することができている．

6 おわりに

本稿では，ツイート中の単語や，気象情報など，ほぼリアルタイムに利用できるデータから風邪の流行を推測する方法について述べた．変数選択では，選択する単語どうしの相関を考慮に入れることで，より風邪流行の特徴を捉えた推測が可能になる．実験では，ビームサーチと mRMR に基づいた選択により，少ない単語数で最も高い推測精度が得られることを示した．

今回は，風邪の流行について推測を試みたが，提案手法自体はその他の事象にも広く適用できる．例えば，商品の売れ行き，株価，世論の動向などの推測や予測にも用いることができる可能性がある．

参考文献

[1] T Sakaki, M Okazaki, Y Matsuo, Earthquake shakes Twitter users: real-time event detection by social sensors, Proceedings of the 19th international conference on World Wide Web (WWW), 2010.

[2] J Bollen and H Mao, Twitter mood predicts the stock market, Journal of Computational Science Vol.2 (1), 2011.

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

[4] H Achrekar, A Gandhe, R Lazarus, Ssu-Hsin Yu and Benyuan Liu, Predicting flu trends using Twitter data, IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2011.

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

[6] PM Polgreen, Y Chen, DM Pennock and FD Nelson, Using Internet searches for influenza surveillance, Clinical Infectious Diseases Vol.47 (11), 2008.

[7] D Das, K Metzger, R Heffernan, S Balter, D Weiss and F Mostashari, Monitoring over-the-counter medication sales for early detection of disease outbreaks – New York City, Morbidity and Mortality Weekly Report Vol.54 (supplement), 2005.

[8] 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.

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

[10] N Kwak and Chong-Ho Choi, Input feature selection for classification problems, Neural Networks Vol.13 (1), 2002.

表 1: 貪欲法での結果 (訓練データとの相関係数，テストデータとの相関係数，選択された単語)

評価式	λ	反復回数	訓練 R	テスト R	選択された単語
人手	-	-	0.832	0.734	風邪
Google	-	-	0.896	0.320	ませ, 了解, たぶん
訓練 R 最大	any	1	0.886	0.331	ませ
mRMR'	0.6	2	0.920	0.453	ませ, 時々
	0.6	3	0.931	0.678	ませ, 時々, 冷やし
	0.6	4	0.937	0.391	ませ, 時々, 冷やし, ゆっくり
	0.6	5	0.943	0.223	ませ, 時々, 冷やし, ゆっくり, He
	0.5	6	0.950	0.205	ませ, フォロー, 多分, 都, ゆっくり, 枝豆
	0.5	7	0.954	0.378	ませ, フォロー, 多分, 都, ゆっくり, 枝豆, 之
MMR	0.6	2	0.920	0.453	ませ, 時々
	0.6	3	0.934	0.624	ませ, 時々, 夏
	0.5	4	0.945	0.649	ませ, フォロー, 晴, 年始
	0.5	5	0.954	0.698	ませ, フォロー, 晴, 年始, 蝉
	0.5	6	0.957	0.721	ませ, フォロー, 晴, 年始, 蝉, 姉さん
	0.5	7	0.960	**0.747**	ませ, フォロー, 晴, 年始, 蝉, 姉さん, 肉
MIFS-U	1.0	2	0.920	0.379	ませ, 晴
	1.0	3	0.933	0.376	ませ, 晴, 広場
	1.0	4	0.940	0.369	ませ, 晴, 広場, 韻
	1.0	5	0.943	0.408	ませ, 晴, 広場, 韻, 加速
	1.0	6	0.945	0.264	ませ, 晴, 広場, 韻, 加速, ワンマン
	1.0	7	0.946	0.250	ませ, 晴, 広場, 韻, 加速, ワンマン, キッチン

表 2: ビームサーチでの結果 (訓練 R 最大値)

評価式	λ	反復回数	訓練 R	テスト R	選択された単語
mRMR'	0.7	2	0.927	0.717	要, (最低気温)
	0.6	3	0.944	0.880	と, 年始, 冷やし
	0.6	4	0.956	**0.894**	と, 年始, 夏, 白菜
	0.6	5	0.962	0.883	と, 年始, 夏, 白菜, He
	0.6	6	0.966	0.853	と, 年始, 夏, 白菜, ゆっくり, He
	0.6	7	0.967	0.823	と, 年始, 夏, 白菜, ゆっくり, He, 有
MMR	0.7	2	0.927	0.717	要, (最低気温)
	0.5	3	0.945	0.744	思い, イレブン, 年始
	0.5	4	0.949	0.796	思い, 晴, 年始, ボディ
	0.5	5	0.955	0.819	思い, たった, 年始, ボディ, 汗
	0.5	6	0.957	0.781	思い, たった, 年始, 練乳, 衛星, ボディ
	0.5	7	0.959	0.745	思い, たった, 年始, 練乳, 衛星, ボディ, 公式
MIFS-U	1.0	2	0.921	0.872	と, 年始
	1.0	3	0.940	0.804	思い, 晴, 末
	1.0	4	0.937	0.864	と, 来年, 吉, 冷却
	1.0	5	0.940	0.852	と, 大掃除, 字, 巫女, 冷却
	1.0	6	0.945	0.841	と, 大掃除, 字, 不安定, 巫女, 冷却
	1.0	7	0.949	0.842	と, 大掃除, 字, 不安定, 巫女, 洞窟, 冷却