

Twitterによる風邪流行の推測



○谷田和章, 荒牧英治, 佐藤一誠, 吉田稔, 中川裕志

東京大学

背景と目的

最近、
風邪が流行ってるし
気をつけようかな…

□ 背景

風邪流行を即座に把握したい

ブログ等への大量の投稿

□ 目的

分析・推測

マイクロブログ: Twitter

□ いまどうしてる?

140 ツイート

□ リアルタイム情報発信

□ ユーザたくさん (ユーザ数>1億, 発言数>2.5億/日)



風邪

- 呼吸器系の炎症の病気の**総称**
- ウイルス**感染**によって発症

風邪が流行

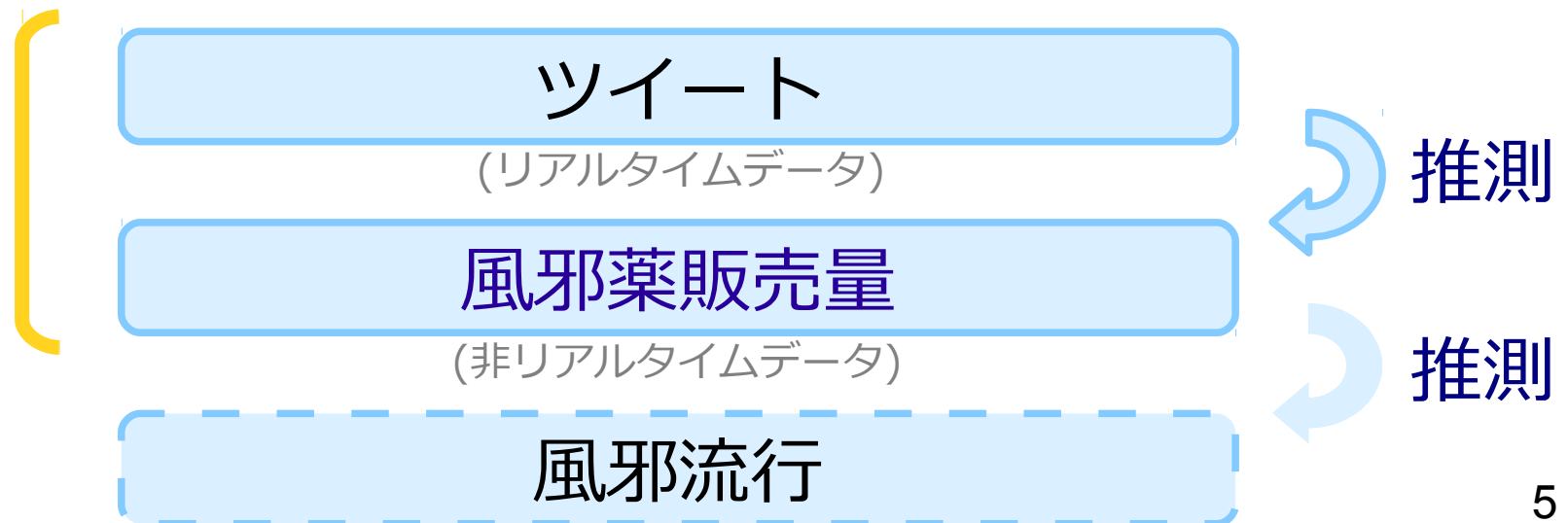


罹患リスクが増大

- 公的な流行の調査はない

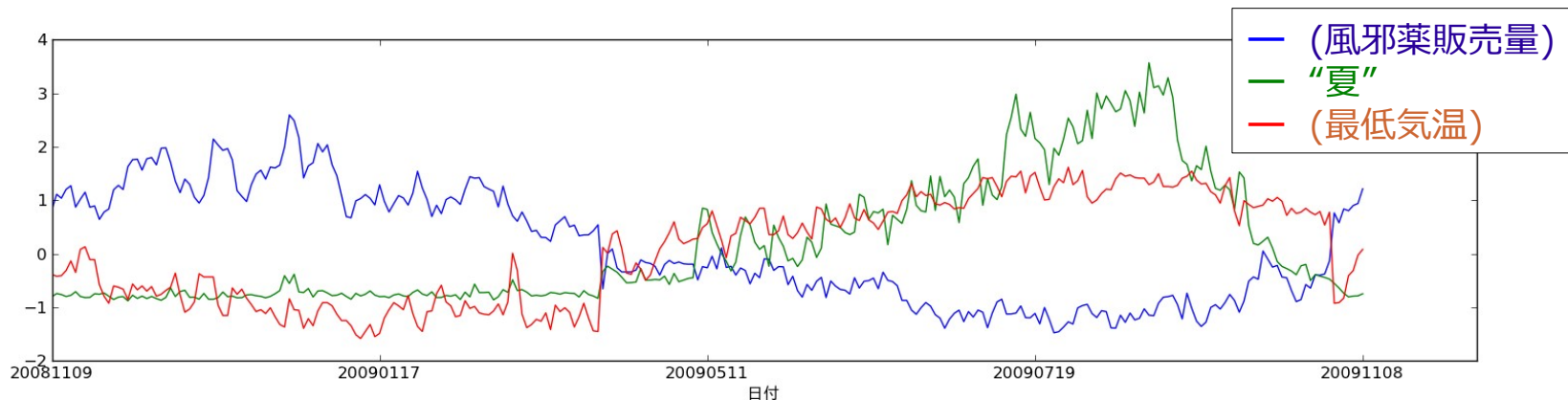
風邪流行の推測

- 薬局の風邪薬販売量
 - 高い相関 (Magruder, 2003)
 - 即時性がない
 - 総務省が翌月下旬に公開
- インターネットを利用



利用データ

- 目的データ (非即時的)
 - 風邪薬販売量
- 説明データ (即時的)
 - ツイート (各単語の出現頻度)
 - 気象情報



ベースライン① <選択, 推測> : 先験的な単語からの推測

□ 単語を人手で選択

<例> x : “風邪”

□ 単語の頻度



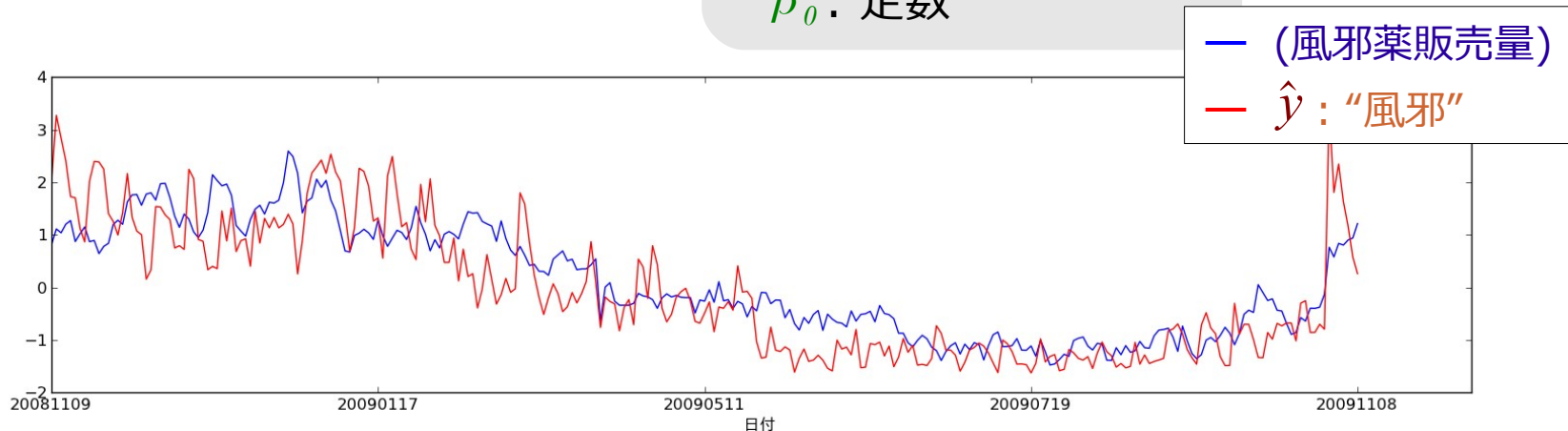
風邪薬販売量

□ $\hat{y} = \beta_0 + \beta_1 x$

x : 単語の出現頻度

β_1 : 重み係数


β_0 : 定数



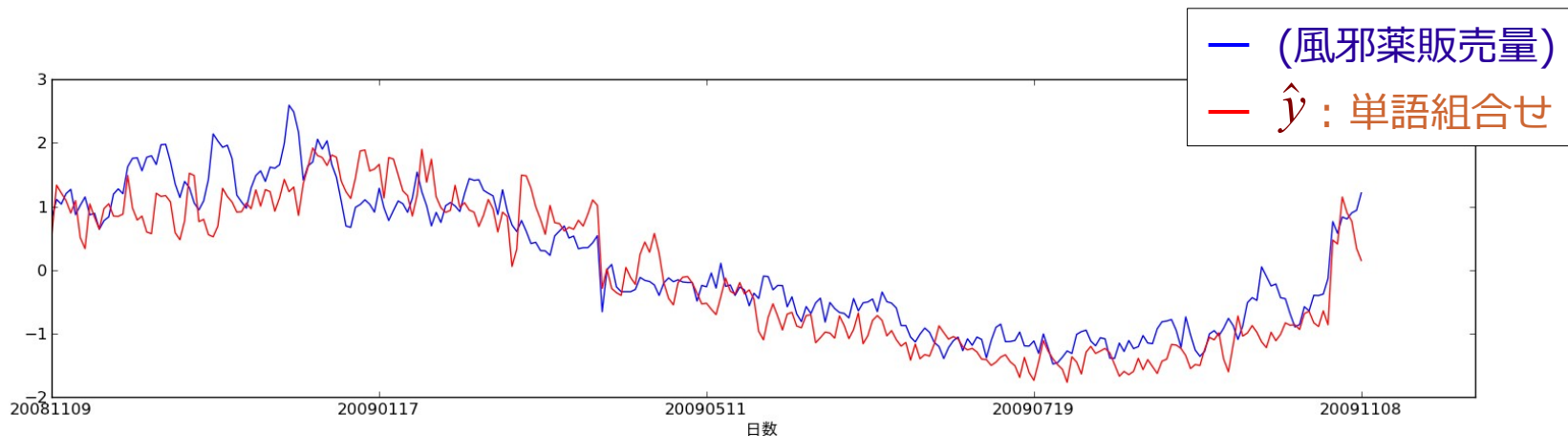
提案手法 <推測> : 重回帰による推測

- 複数の単語を選択

<人手による例>
"風邪", "寒い", (最低気温)

- 複数の単語  風邪薬販売量
推測

- $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$



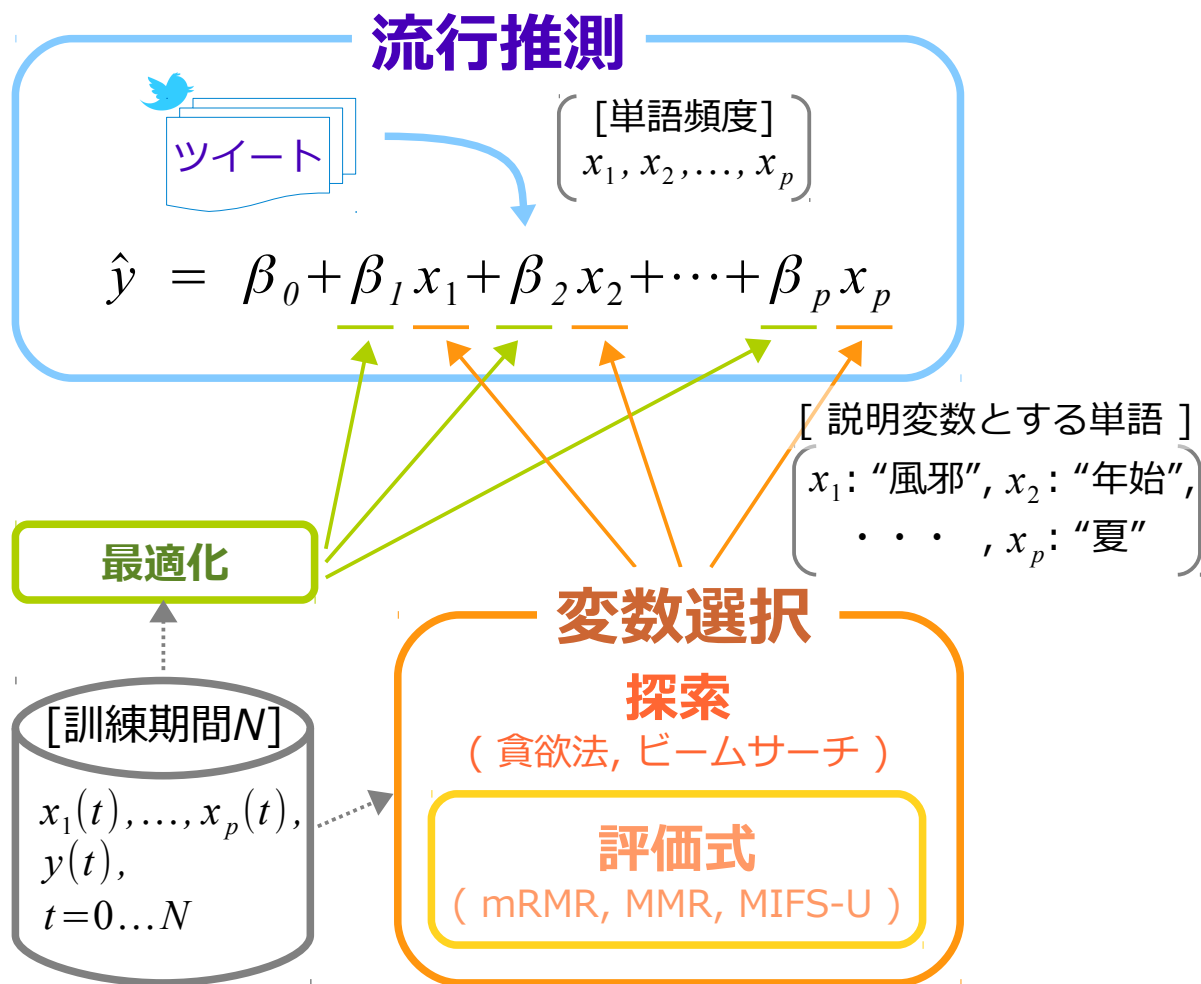
提案手法の全体像

1. 変数選択

- 評価式
- 探索法

2. 最適化

3. 流行推測



提案手法： 変数選択

変数選択

探索

(貪欲法, ビームサーチ)

評価式

(mRMR, MMR, MIFS-U)

- 説明変数とする単語を複数選ぶ
- 目的
 - 風邪薬販売量と相関の強い単語集合を得る
 - しかし、全ての組合せで最適化して相関を計算するのは困難
- 手法
 - 評価式 (mRMR, MMR)
 - 最適化の代わりに計算
 - 計算コストが比較的少ない
 - それでも全ての組合せの計算は困難
 - 探索 (貪欲法, ビームサーチ)
 - 評価する組合せを限定

提案手法： 評価式

- ある単語の説明変数としての良さ
 - 冗長性を考慮した素性選択法を応用

- mRMR' (minimum Redundancy Maximum Relevance)

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

風邪薬販売量 選択済み単語との類似度
との類似度

λ : 重み

R : 相関係数

S : 選択済み集合

※ 時系列データ ※

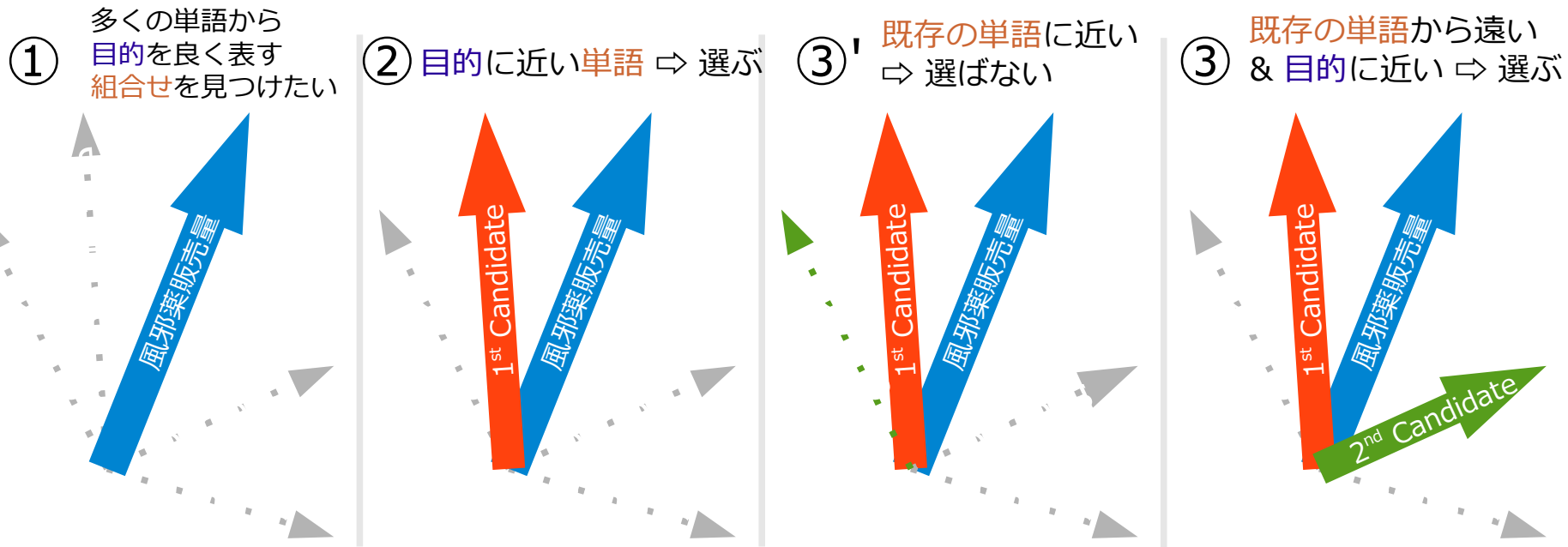
y : 風邪薬販売量

x : 単語頻度

- MMR (Maximal Marginal Relevance)

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

提案手法 <探索> : 貪欲法による変数選択



□ 貪欲法

1. 評価式を最大にする単語 x を選択
2. 単語 x を選択済み集合に加える
3. 1.から繰返し

提案手法 <探索> : ビームサーチによる変数選択

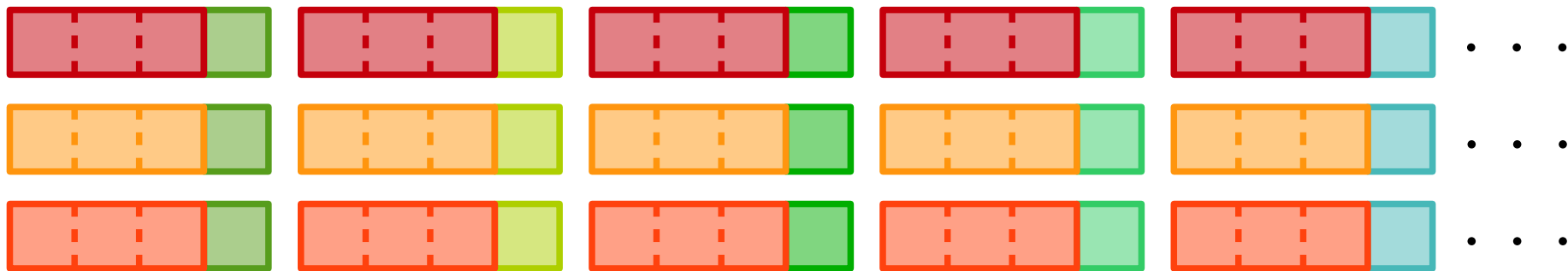
選択済み単語集合の集合

ビーム幅=3



全単語

各単語集合に単語を付ける



評価が上位の単語集合



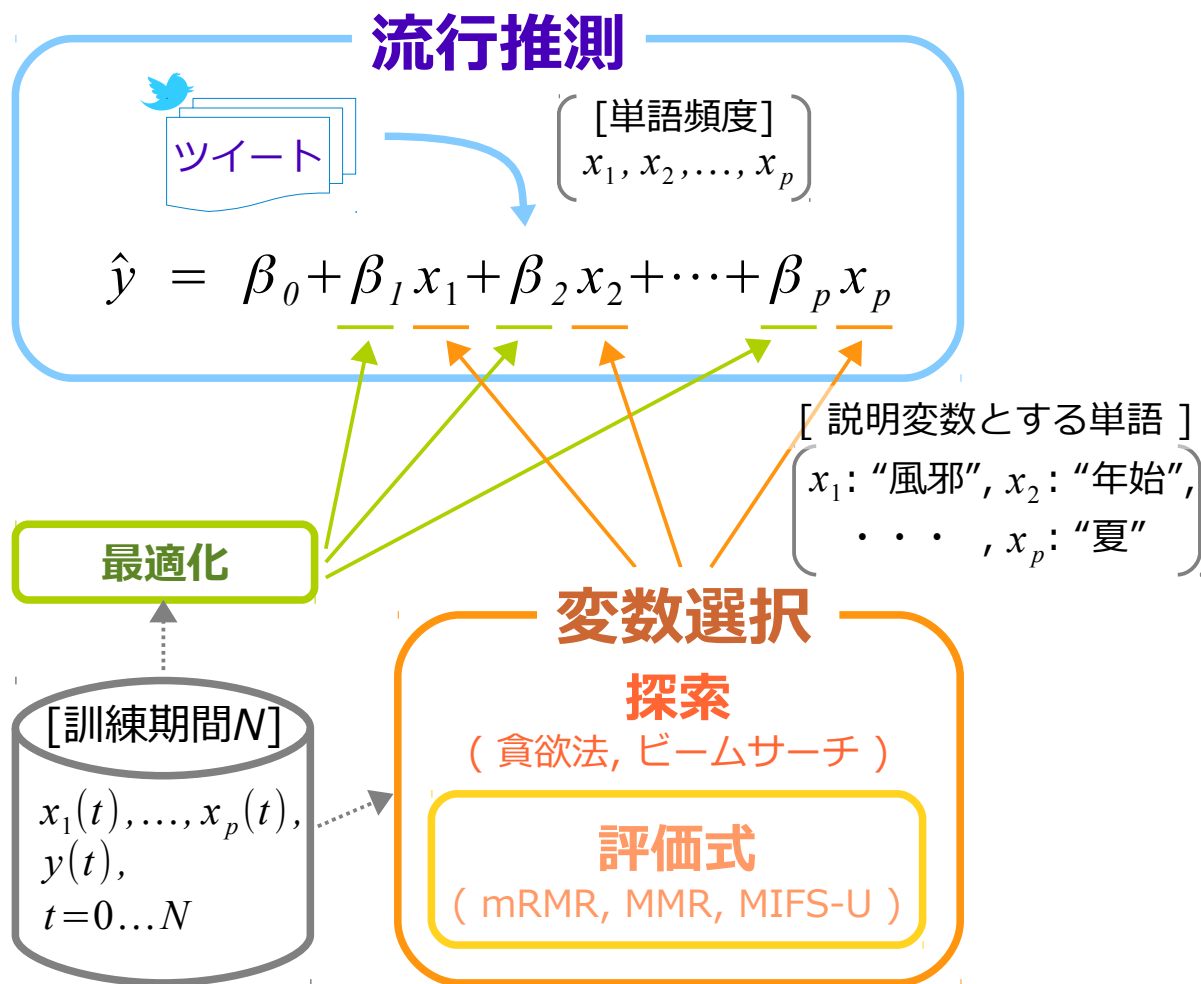
提案手法の全体像 (再掲)

1. 変数選択

- 評価式
- 探索法

2. 最適化

3. 流行推測



実験: 結果

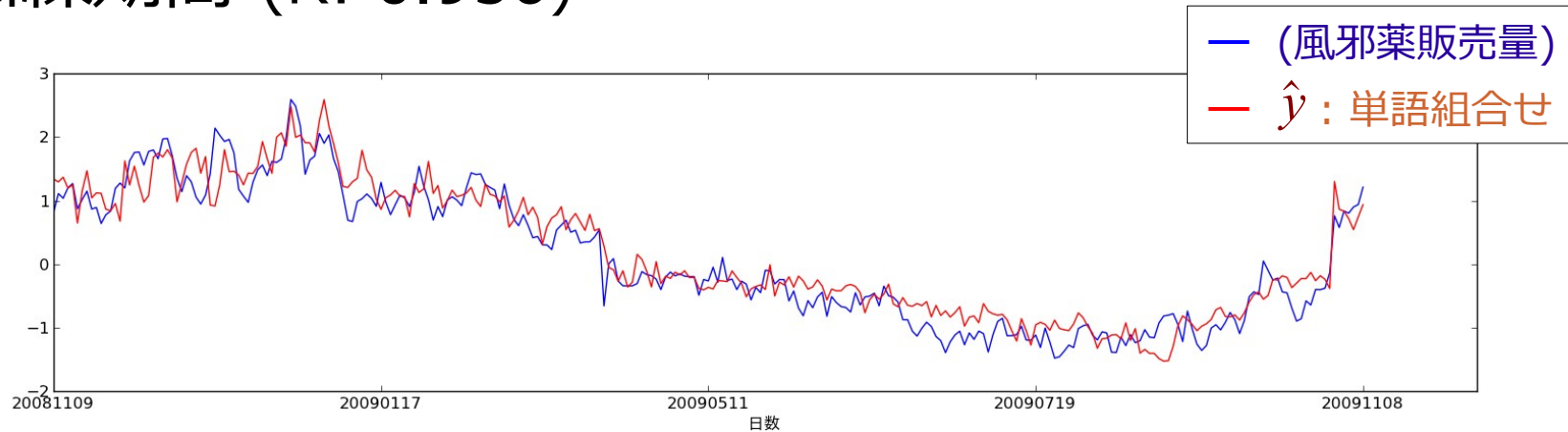
- 訓練: 2008/11/09 – 2009/11/08
- テスト: 2009/11/09 – 2010/07/04
2011/06/05 – 2010/08/31
- 探索: $\lambda = 0.1$ to 0.9 step 0.1 , 単語数 = 1 to 7

方法	λ	単語数	訓練R	テストR
人手 (“風邪”)	-	1	0.832	0.734
Google (Ginsberg 2009)	-	(3)	0.896	0.320
訓練R最大	any	1	0.886	0.331
貪欲法 : MMR	0.5	7	0.960	0.747
ビームサーチ : mRMR'	0.6	4	0.956	0.894

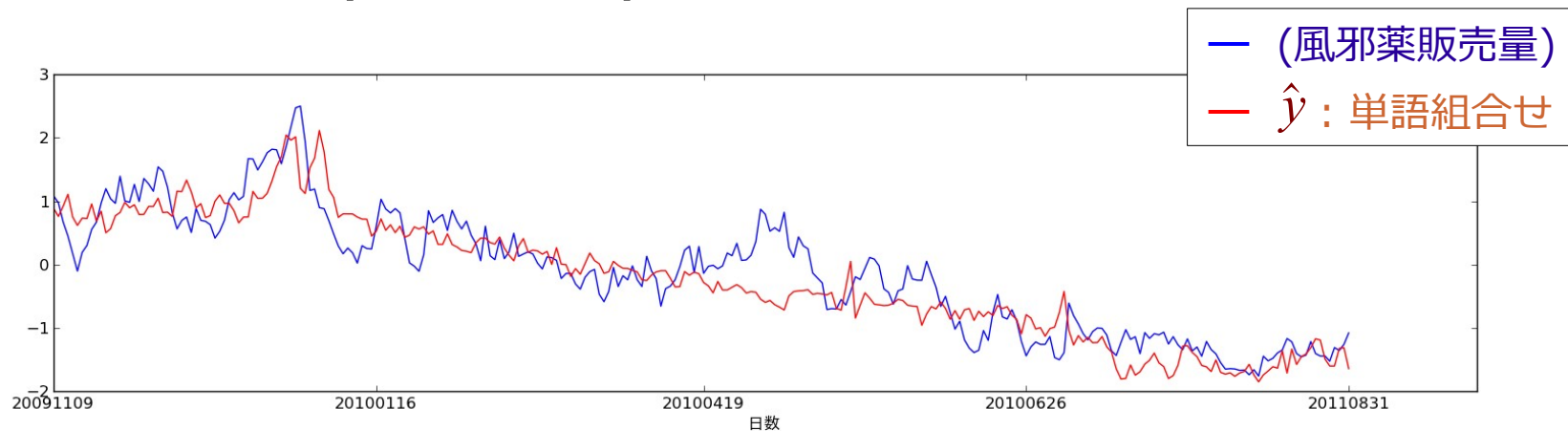
※ 貪欲法:MMRは、貪欲法のなかで訓練R最大。ビームサーチ:mRMRは、単語数4で訓練R最大

実験 (ビームサーチ:mRMR', $\lambda=0.6$, 単語数=4) : “と”, “年始”, “夏”, “白菜”

□ 訓練期間 (R: 0.956)

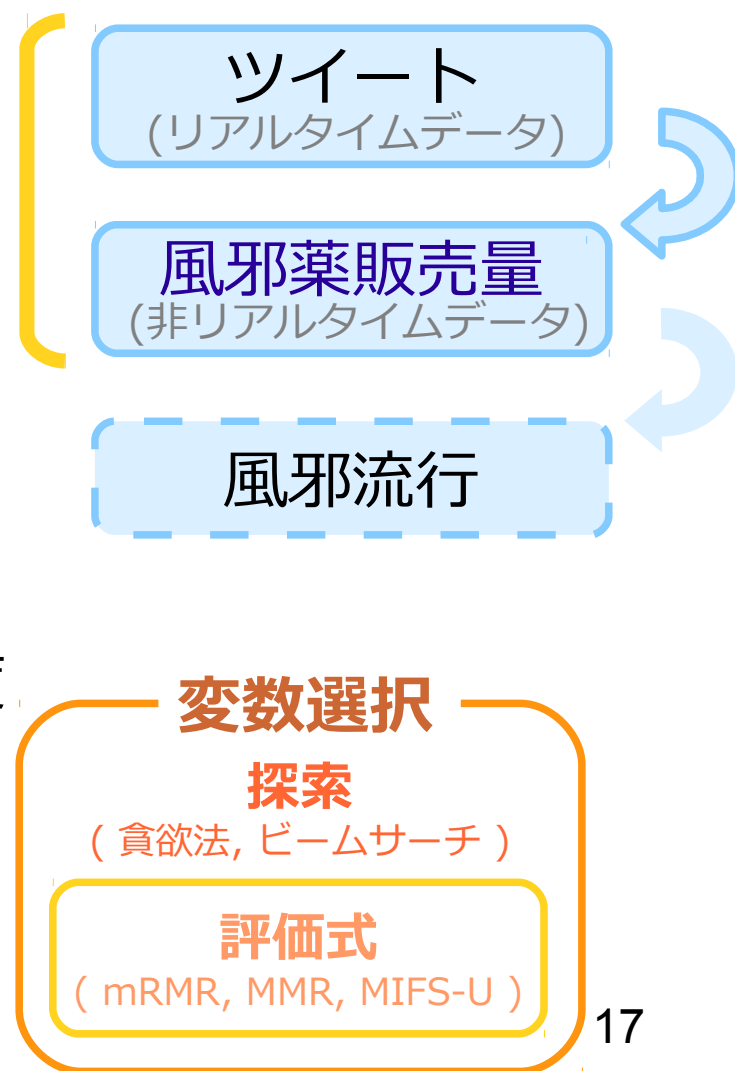


□ テスト期間 (R: 0.894)



まとめ

- 風邪流行: 風邪薬販売量から推測
- 風邪薬販売量: ツイートから推測
- 単語の選択
 - 評価式: mRMR
 - 探索法: ビームサーチ
 - 少ない単語数でも高い推測精度
 - 少ない計算コスト
- 結果: テスト期間R:0.894
("風邪" R:0.734)



おまけ：

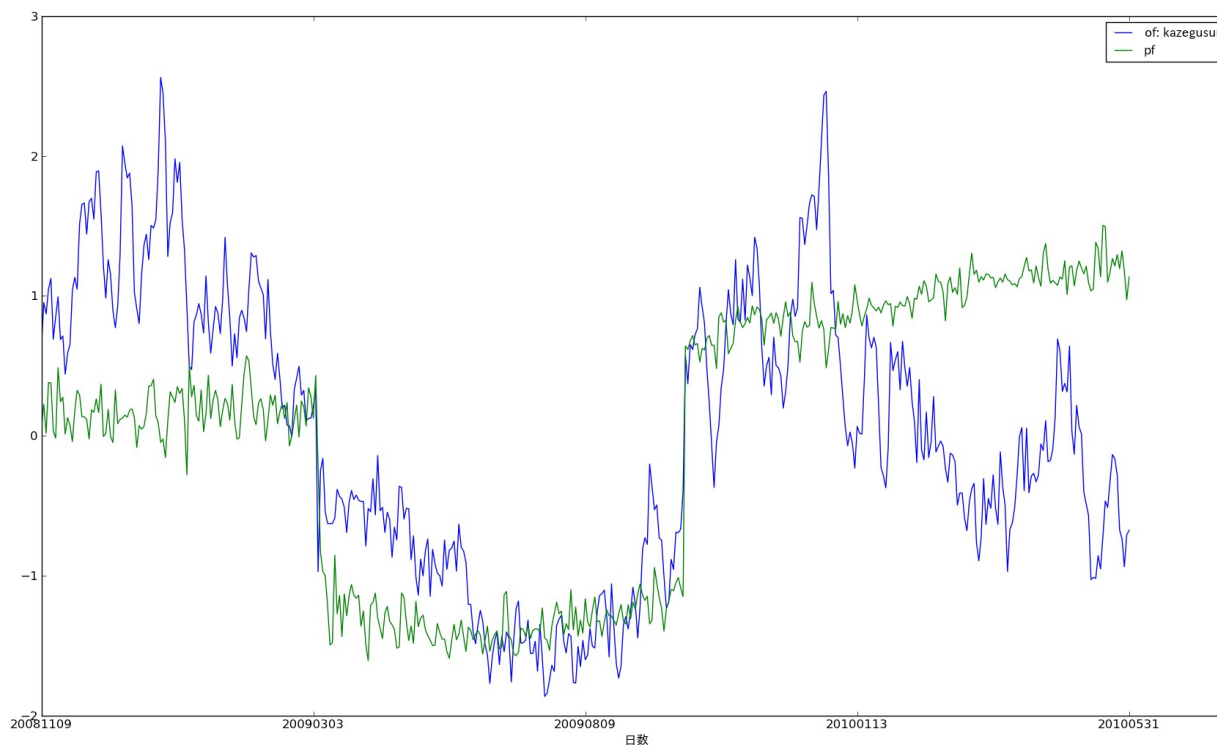
直近の風邪流行を推測してみた

參考資料

- 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.
- J Bollen and H Mao, Twitter mood predicts the stock market, Journal of Computational Science Vol.2 (1), 2011.
- 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.
- 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.
- 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.
- PM Polgreen, Y Chen, DM Pennock and FD Nelson, Using Internet searches for influenza surveillance, Clinical Infectious Diseases Vol.47 (11), 2008.
- 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.
- 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.
- J Carbonell and J Goldstein, The use of MMR, diversity-based reranking for reordering documents and producing summaries, Proceedings of the SIGIR, 1998.
- N Kwak and Chong-Ho Choi, Input feature selection for classification problems, Neural Networks Vol.13 (1), 2002.

“ませ” training: 08'11'09-09'11'08

- 訓練: 0.886
- テスト: -0.638



実験: 結果

- 訓練: 2008/11/09 – 2009/11/08
- テスト: 2009/11/09 – 2010/07/04
2011/06/05 – 2010/08/31
- 探索: $\lambda = 0.1$ to 0.9 step 0.1 , 単語数 = 1 to 7

方法	λ	単語数	訓練 R	テスト R	選択された単語
人手	-	1	0.832	0.734	“風邪”
Google (Ginsberg)	-	(3)	0.896	0.320	“ませ”, “了解”, “たぶん”
訓練R最大	any	1	0.886	0.331	“ませ”
貪欲法:MMR	0.5	7	0.960	0.747	“ませ”, “フォロー”, “晴”, ...
ビームサーチ:mRMR'	0.6	4	0.956	0.894	“と”, “年始”, “夏”, “白菜”

※ 貪欲法:MMRは、貪欲法のなかで訓練R最大。ビームサーチ:mRMRは、単語数4で訓練R最大 21