# Twitter Listを用いた低コストな Tweet分類器の学習

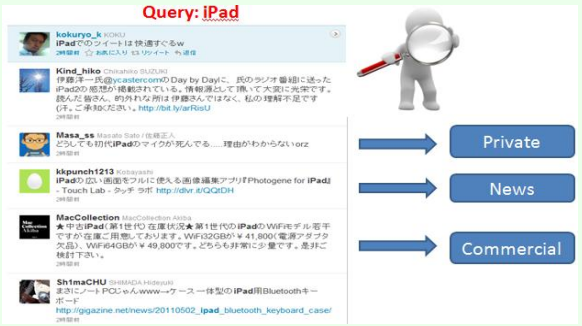
鄭 洪光, 吉永 直樹, 鍛治 伸裕,豊田 正史

東京大学 情報理工学系研究科 喜連川・豊田研究室

## ■研究背景

- 近年、Microblog、特にTwitterは(災害情報やイベント中継など)重要なリアルタイム情報収集のための情報源として、驚異的な成長を見せている
- 情報収集支援を目的として、Twitter に日々投稿される膨大なTweetを、ユーザの検索ニーズに合わせたカテゴリー(ニュース、広告、個人呟き)に分類することを目指す

分類器を構築するための大規模な学習データをどのように集めるか？




## ■研究概要

- Tweetカテゴリーの定義
  - Private: 個人経験や意見
  - Commercial: 宣伝
  - News: 客観的なニュース


カテゴリー間の違いに基づき、二種の手法を提案する

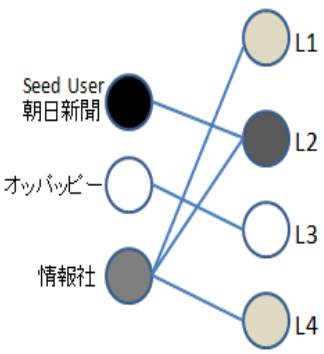
- 提案手法①:ユーザのリストを用いたラベル伝搬による学習データの収集 ( News & Commercial )

  News seed users: @googlenewsjp, @nikkeitter, @47news, @yomiuri_online, @YahooNews, @asahi …

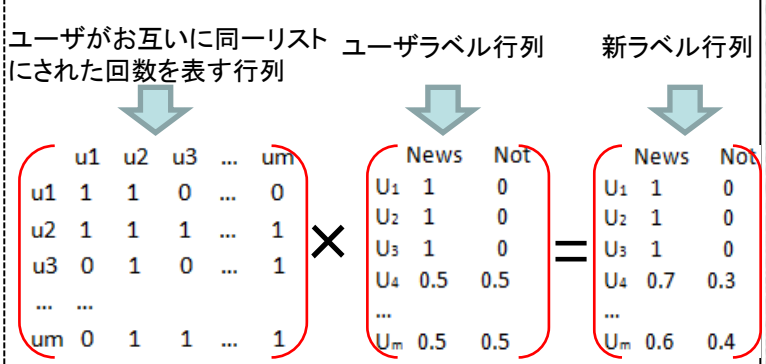
  Commercial seed users: @yahoo_shopping, @kadenbest, @ranranraku,@kaimonosuki, @yellclick …

◆ 1. User-List


2. ラベル伝搬のイメージ




3. ラベル伝搬によるユーザラベルの計算




- 提案手法②:ユーザのプロフィール情報を用いたアカウントの収集 ( Private )

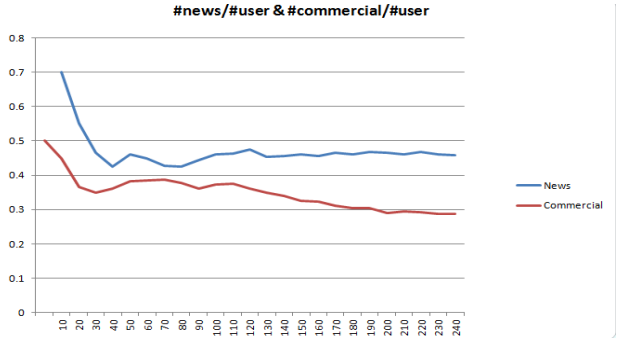
  ◆ アカウントのユーザ名を形態素解析し、人名であると判定されたアカウントからtweetを自動収集

  例:「優さん」→ 優 名詞,固有名詞,人名,名,*,*,優,ユウ,ユー ; さん 名詞,接尾,人名,*,*,*,さん,サン,サン

  「福間健二」→ 福間 名詞,固有名詞,人名,姓,*,*,福間,フクマ,フクマ ; 健二 名詞,固有名詞,人名,名,*,*,健二,ケンジ,ケンジ

## ■実験

■ラベル伝搬によるユーザアカウントの精度




■実験設定

| | News | Commercial | Private |
|---|---|---|---|
| Tweet数 (アカウント数) | 23,683 (100) | 20,068 (98) | 23,263 (196) |

- 分類器: SVM

  特徴量:bag-of-words (BOW);

  #friends #followers ; domain of url

  5分割交差検定

## ■実験結果

| 特徴量 | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| BOW | 0.828 | 0.817 | 0.817 | 0.817 |
| + ln(#friends) + ln(#followers) | 0.875 | 0.873 | 0.873 | 0.873 |
| + ln(#friends) + ln(#followers)+url_domain | 0.827 | 0.831 | 0.842 | 0.836 |

## ■今後の課題:

- 実験結果の分析 & 一般性を持つテストデータの構築及び新しい特徴量の検討