典型的情報発信者に着目したMicroblog分類器学習の低コスト化に関する一考察

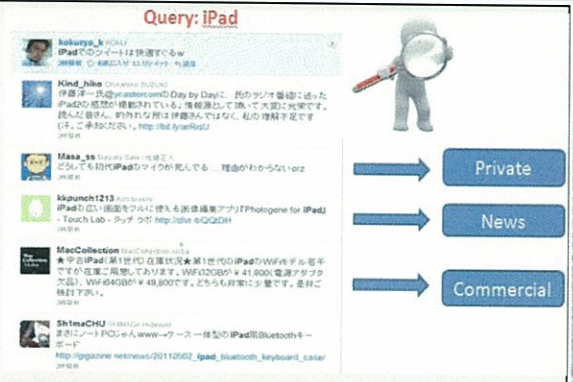
鄭 洪光(テイ コウコウ), 鍛治 伸裕, 吉永 直樹, 豊田 正史

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、異なる意図を持って発信されたtweetを、ユーザの検索ニーズに合わせたカテゴリー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
  - News: 客観的なニュース
  - Commercial: 宣伝
- 分類器: SVM

  特徴量:bag-of-words (BOW); #friends #followers
- 提案手法:典型的情報発信者に着目したデータの収集

  各カテゴリー毎に典型的な情報発信者が存在することに着目し、情報発信者にラベル付けすることで大規模のラベルデータを収集する
  - ◆ 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
  - ◆ News & Commercial: 少数のアカウントからtweetを人手で収集

    Newsアカウント: @asahi, @mainichijpedit, @yomiuri_onlie, @YahooNewsTopics, @livedoornews …

    Commercialアカウント: @mixprice_com, @rakuraku360, @kaimonosuki, @ranranraku, @yellclick …

■実験設定

	News	Commercial	Private
Tweet数 (アカウント数)	38,441 (10)	50,580 (10)	62,667 (12,533)

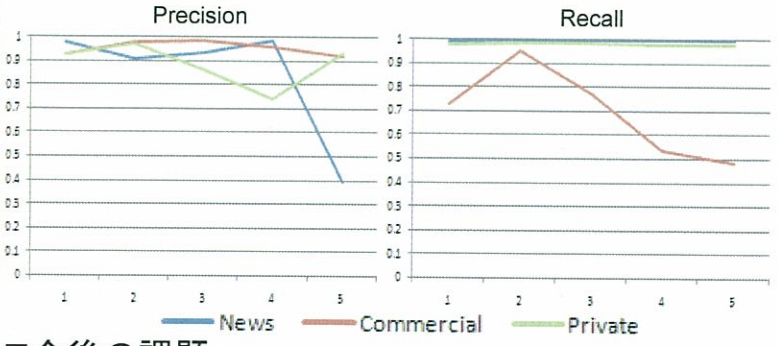
●5分割交差検定

学習データとテストデータで発言期間とユーザアカウントが重複しないように分割

※各Privateアカウントから5tweetを自動収集

■実験結果

特徴量	Accuracy	Precision	Recall	F1
BOW	0.859	0.784	0.841	0.812
+ ln(#friends) + ln(#followers)	0.901	0.893	0.891	0.892



■各カテゴリーにおける単語の重み

News	重み	Commercial	重み	Private	重み
rank	2.19	セール	1.48	@	1.14
共同	1.65	|	1.28	笑	1.02
インタビュー	1.36	くよくよ	1.23	#	0.93
無明	1.09	マイシティーサーチ	1.17	w	0.87
ヨミ	1.01	楽天	1.05	www	0.80
Sports	0.81	――――	1.00	眠い	0.79
ドアスポ・ドアブロ	0.80	ネイティブ	0.98	―	0.75
読売新聞	0.79	やさしい	0.94	ww	0.74
ドクター	0.78	ひと休み	0.94	しかし	0.73
Watch	0.74	大した	0.90	マジ	0.71

■今後の課題:

- 実験結果の分析
- 一般性を持つテストデータ及び新しい特徴量の検討