

大規模ウェブデータからの半教師あり意味知識獲得

小町研究室（自然言語処理） 准教授 小町守 / 協力: Microsoft Research 鈴木久美, ヤフー株式会社 内海慶・颯々野学

概要

少数の意味カテゴリつき固有表現を大規模なウェブ検索ログと組み合わせることで、高精度に意味カテゴリつき固有表現を抽出することができる。

- ① 形態素解析（単語分かち書き）を用いず、頑健にウェブデータを処理する手法を開発した
- ② 検索クエリログ・検索クリックスルーログの自然言語処理タスクへの有効性を示した
- ③ Google の並列分散処理環境 MapReduce を用いたスケーラブルなアルゴリズムを提案した

背景

- ✓ウェブユーザの検索意図を知りたい
- ✓意味カテゴリのついた単語がほしい

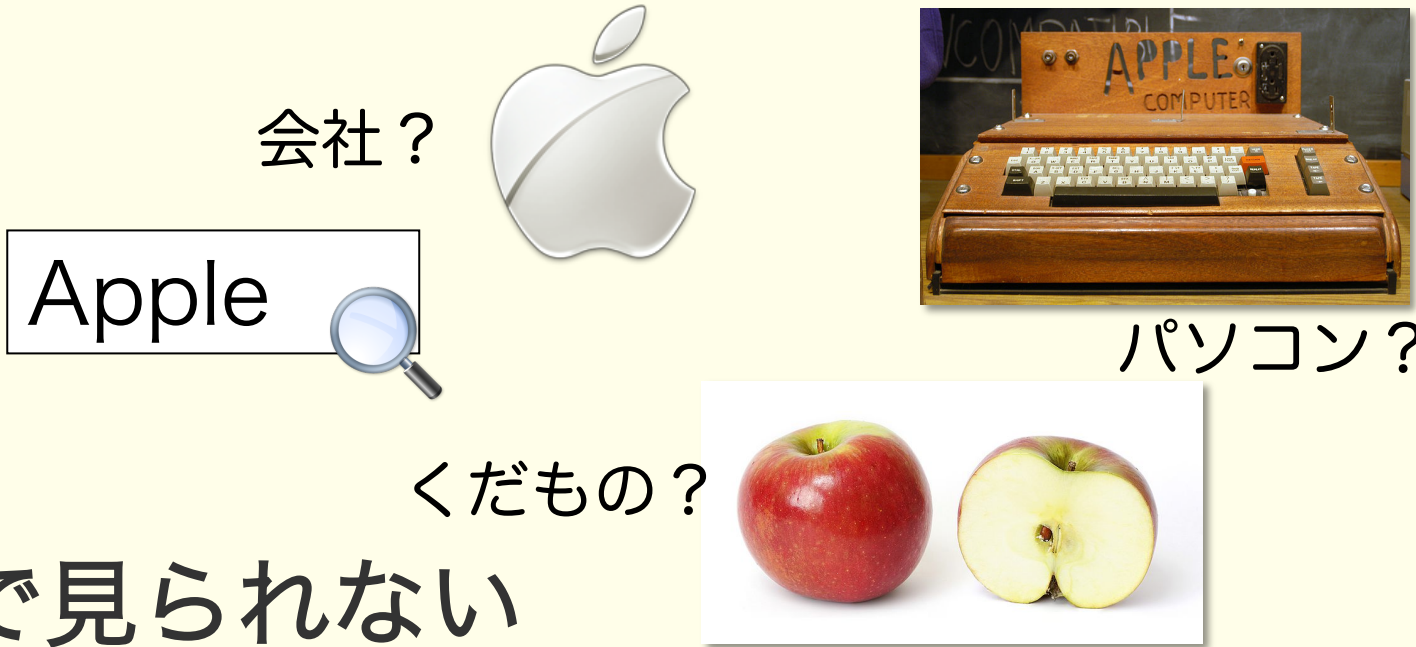
問題点

- ①データが大量にあるので全部を手で見られない
- ②ウェブデータは碎けた表現や書き誤りが多いので解析しにくい

目的

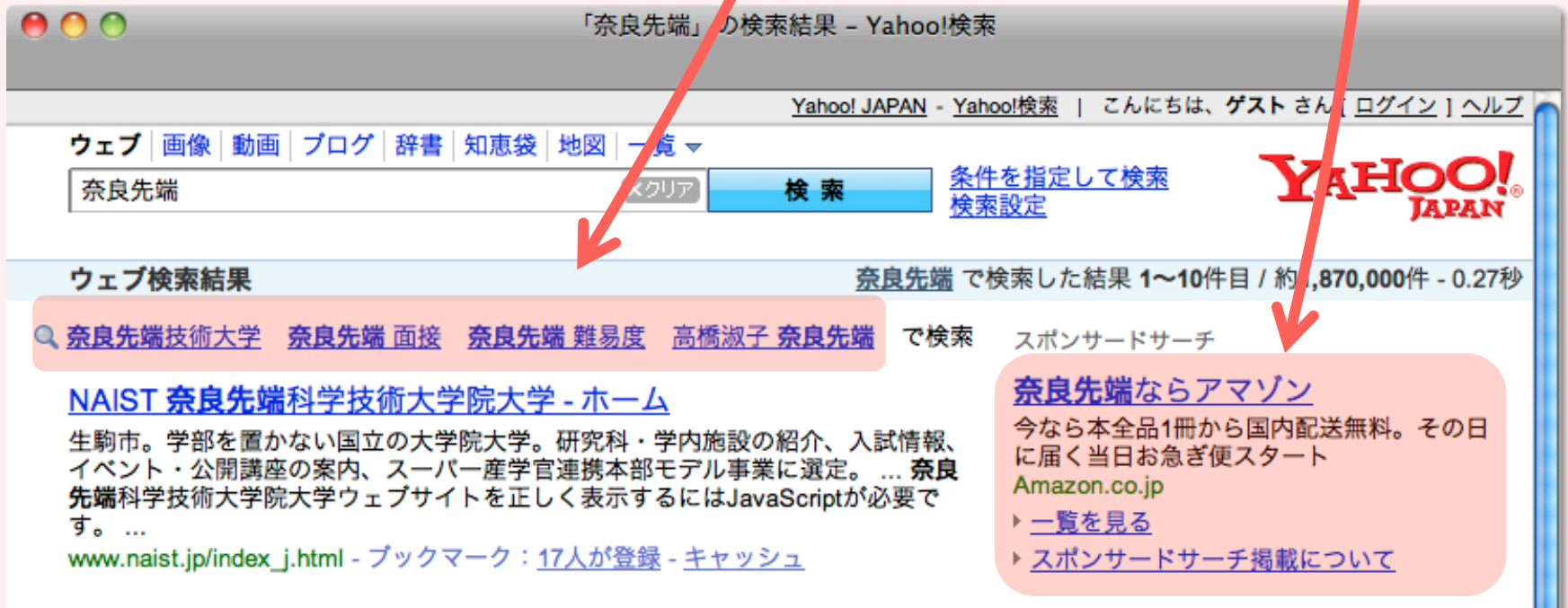
- ①少数のデータから大規模に学習するアルゴリズムの開発
- ②言語非依存の頑健な解析手法の開発

研究背景



応用

- ✓検索の連動広告
- ✓関連検索
- ✓情報推薦



提案手法

ラプラシアンを用いたラベル伝播
MapReduce (Hadoop) による並列分散計算

アルゴリズム

入力:

- シードインスタンス行列 $\mathbf{F}(0)$
- インスタンス類似度行列 $\mathbf{A}$

出力:

- インスタンススコア行列 $\mathbf{F}(t)$

1.正則化ラプラシアン行列を作成

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}$$

2.収束するまで以下を繰り返す

$$\mathbf{F}(t+1) = \alpha(-\mathbf{L})\mathbf{F}(t) + (1 - \alpha)\mathbf{F}(0)$$

最終的なラベルcは以下で決定

$$\hat{c} = \underset{c}{\operatorname{argmax}} \mathbf{F}_c(t)$$

$$\mathbf{A} = \mathbf{W}^T \mathbf{W}$$
$$\mathbf{W} = \frac{|x_i, p_j|}{\sum_k |x_i, p_k|}$$
$$\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$$

(次数対角行列)

関連研究

- ①巨大な類似度行列を疎な重み行列Wで保持し、記憶容量を削減
 - ✓ $O(n^2) \rightarrow O(np')$ (ただし n はインスタンス数、p'は平均パターン数)
- ②アルゴリズムの反復回数を定数にすることで行列の固有値計算を近似
 - ✓ $O(n^3) \rightarrow O(np't)$ (ただし t は反復ステップ数)
 - ✓ MapReduce によって $O(np')$ の計算は並列分散化可能

名前	シード数	抽出対象	反復回数	コーパス	言語
Tchai	5	意味カテゴリつき単語	∞	検索クエリログ	日本語
Quetchup	5	意味カテゴリつき単語	∞	検索クエリログ+ 検索クリックスルーログ	日本語
Sekine & Suzuki	~600	カテゴリつき固有表現	1	検索クエリログ	英語
Basilisk	10	意味的語彙	∞	MUC-4 [Sundheim 92]	英語
Espresso	~10	意味的關係	∞	TREC09 [Voorhees 01]	英語

実験データ

Yahoo! 検索2008年8月の検索ログ集合のうち異なり頻度上位1,000万クエリ (数十GB)

クリックスルーパターン

クエリに対してクリックされたアドレス

(例) 「ANA」 → 「www.ana.co.jp」

クエリパターン

空白で区切られた2単語クエリから得られる文脈パターン

(例) 「JR 時刻表」 → 「# 時刻表」

意味カテゴリ

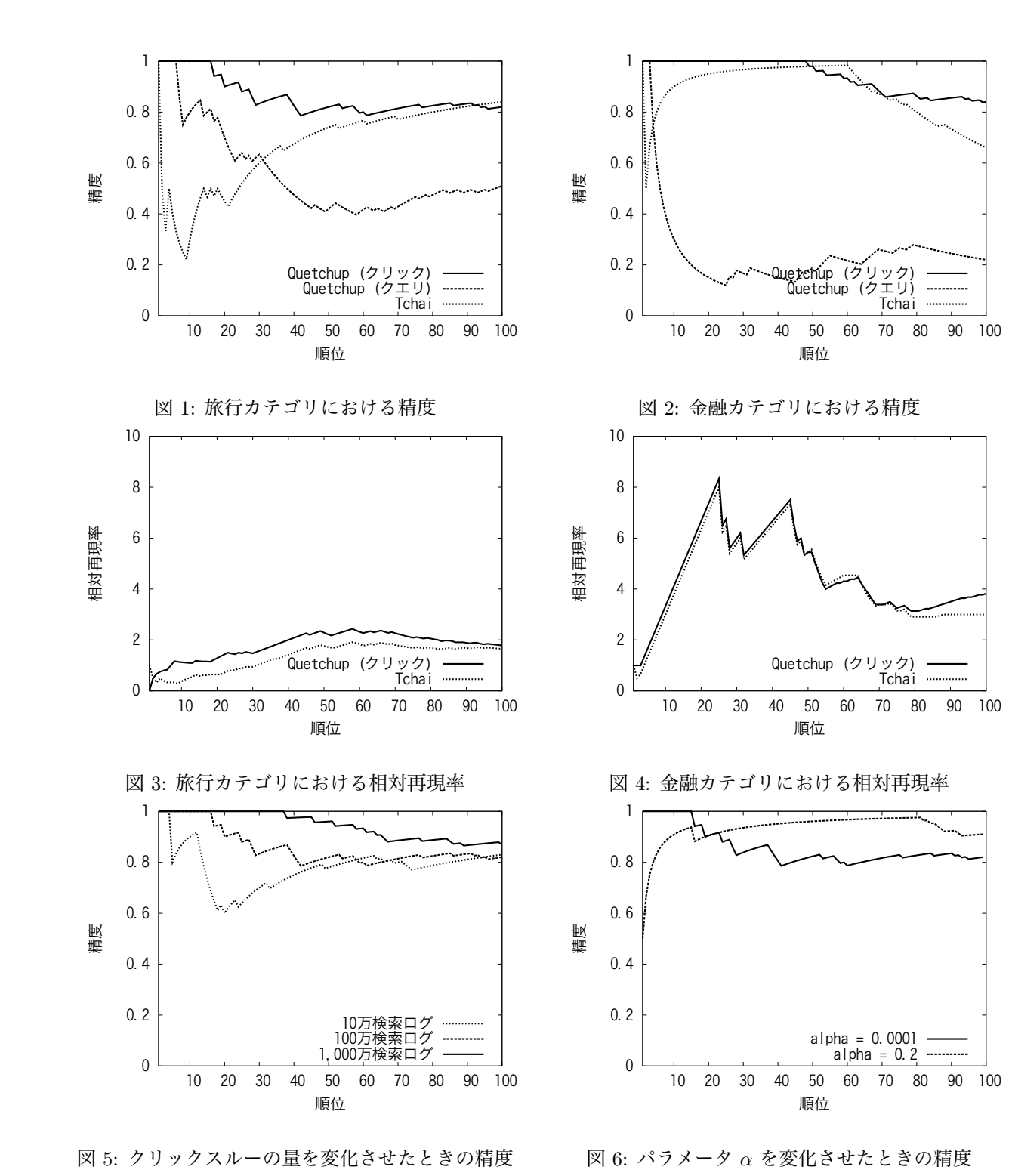
カテゴリ	シード
旅行	jal, ana, jr, じゃらん, his
金融	みずほ銀行, 三井住友銀行, jcb, 新生銀行, 野村證券

評価尺度

- ✓順位kまでの精度 (precision at k)
- ✓相対再現率 [Pantel 04]

実験

実験結果



タイプ	数	例
交通機関	54	広島 新幹線, 東海道線, jr 飯田線, jr 博多, 京都 新幹線
宿泊施設	10	ホテルビナス, リーガロイヤルホテル大阪, ホテル京阪ユニバーサル・シティ, 札幌全日空ホテル, www.route-inn.co.jp
旅行情報	10	外務省 安全, チケットショップ大阪, 観光 関西, 高山観光協会, グーグルナビ
旅行会社	6	jr おでかけネット, 近畿ツー, タビックス 静岡, フレックス インターナショナル, オリオンツアー
その他旅行	2	プロテカ, jal 紀行倶楽部
それ以外	20	格安航空チケット 海外, 新幹線予約状況, 新幹線 時刻表, 温泉宿, 新幹線 停車駅, 虎, youtube 海外ドラマ, 法務部採用, おくりびと, 社会人野球

(上位10,000件のランダムサンプリング)