クエリログをコーパスとした意味知識獲得法の改善

伊藤 淳† 戸田 浩之‡ 廣嶋 伸章‡ 望月 崇由†† 鈴木 智也†† 筧 捷彦†*

†早稲田大学理工学術院 基幹理工学研究科 情報理工学専攻
〒169-8555 東京都新宿区大久保 3-4-1

‡日本電信電話株式会社 NTT サイバーソリューション研究所
〒239-0847 横須賀市光の丘 1-1

††NTT レゾナント株式会社
〒108-0023 東京都港区芝浦 3-4-1 グランパークタワー

E-mail: †ito-jun@kake.info.waseda.ac.jp, ‡{toda.hiroyuki, hiroshima.nobuaki}@lab.ntt.co.jp,
and †*kakehi@waseda.jp

あらまし 本研究では，Web 検索エンジンのクエリログをコーパスとしたブートストラッピングによる意味知識獲得の改善手法を提案する．意味知識獲得とは，同じカテゴリに属するいくつかのキーワードをシードとして，同じカテゴリに属する新たなキーワード群を抽出する手法である．提案手法はブートストラッピングに特有である意味ドリフトという問題をパターン選択時にフィルタをかけることで回避し，既存手法よりも F 値で 1.26 倍の精度向上を達成することができた．

キーワード 情報抽出，テキストマイニング，Set Expansion，Query Logs

Improvement of the Knowledge Acquisition Method using Query Logs

Jun Ito† Hiroyuki Toda‡ Nobuaki Hiroshima‡ Takayoshi Mochizuki†† Tomoya Suzuki††
and Katsuhiko Kakehi†*

†Department of Computer Science and Engineering, Fundamental Science of Engineering, Waseda University
3-4-1 Okubo, Shinjuku, Tokyo, 169-8555 Japan

‡NTT Cyber Solutions Laboratories, NTT Corporation
1-1 Hikarino-oka, Yokosuka, Kanagawa, 239-0847 Japan

††NTT Rezonant Inc.
4-1-8F Granpark tower, Shibaura 3-chome, Minato-ku, Tokyo, 108-0023 Japan

E-mail: †ito-jun@kake.info.waseda.ac.jp, ‡{toda.hiroyuki, hiroshima.nobuaki}@lab.ntt.co.jp,
and †*kakehi@waseda.jp

Abstract In this paper, we propose a improving bootstrapping-based knowledge acquisition method using query logs. This method is a technique to extract new keywords using some same category keywords as seeds. In order to extract keywords, we have to choose patterns which appear together. But, if we choose too generic patterns, we extract keywords which are not extraction category. This problem is called *Semantic Drift*. We avoided *Semantic Drift* by applying the pattern filtering we proposed. As a result, we were able to achieve precision of 1.26 times in F-measure than the existing method.

Keyword Information Retrieval, Text Mining, Set Expansion, Query Logs

1. はじめに

近年，検索エンジンが提示する検索結果の精度が向上し，ユーザは求める情報を載せた Web ページへ容易にアクセスできるようになった．しかし，従来の検索エンジンは，基本的にユーザが入力したクエリが文字列として含まれている Web ページを検索するものであり，クエリに込められたユーザの意図を理解したり，クエリそのものの意味を理解したりして検索を行っているわけではない．そのため，クエリを単なる文字列として扱うのではなく，クエリが持つ意味やクエリが属するカテゴリなどを考慮することで，より適切な検索結果をユーザへ提示することができると考えられて

おり，各検索エンジン事業者は様々な試みを始めている．例えば，Google[1]では，住所をクエリとして入力すると，Google Maps[2]が検索結果の上位に表示されるようになっている．また，Yahoo![3]では，“天気 東京”と入力すると，東京の天気情報がアイコンつきで検索結果の上位に表示されるようになっている．

各検索エンジン事業者がこのようなサービスをどのように実現しているのかは明らかにされていない．しかし，キーワードとカテゴリの関係を示した辞書を保持しておき，検索時にキーワードのカテゴリを特定して検索結果を切り替えることで，このようなサービスが実現できるのではないかと考えられる．例えば，“東京”は地名カテゴリ，“浜崎あゆみ”はアーティストカテゴリであるということが分かると，天気情報を示したり，アーティスト情報を検索結果の上位に示したりすることが可能になる．

このようにして実現する場合，カテゴリに属するキーワード群をあらかじめ用意しておく必要がある．地名ならば，地理データなどから簡単に用意できるかもしれないが，アーティストや映画，番組など，次々と新しいものが生まれるようなカテゴリでは，キーワード群を常に最新の状態に保つことは手間がかかる．また，カテゴリの数は膨大であり，とてもすべてのカテゴリとキーワードを人手で収集し，分類することはできない．そのうえ，カテゴリの中には専門知識を扱うようなものも存在するため，その分野に精通したエキスパートのチェックが必要になることも考えられる．そのため，なるべく人の手間をかけずに，自動的に，新語にも対応できるような，カテゴリ情報つきキーワード群の抽出手法が重要と考えられる．

そこで本研究では，クエリログをコーパスとした意味知識獲得手法に着目した．クエリログは新しい用語が反映されやすいうえ，クエリ文字列が適切なキーワードごとに分割されている．わかち書きが必要な日本語においては，クエリログをコーパスとして利用することは特に有用であると考えられる．

2. 関連研究

Pantel ら[1]は，Espresso というアルゴリズムを提案した．Espresso は，“X is a Y”など，係り受け関係に着目してキーワードを抽出する．キーワード抽出に用いられる“is a”などのパターンと，抽出された“X”や“Y”などのキーワードは，Pointwise Mutual Information (PMI)を用いた信頼度によってランキングされる．パターンとキーワードの信頼度は，お互いの信頼度を利用して計算するように定義されているため，ブートストラッピング手法になっている．小町ら[2]は，Espresso をクエリログコーパスに適用できるように変更した Tchai というアルゴリズムを提案した．Espresso において問題であった実行速度の遅さを改善したほか，後述する意味ドリフト問題を共起パターンや共起インスタンス数の最大値に着目して改善している．

ブートストラッピングを用いないインスタンス抽出手法としては，Cafarella ら[3,4]が提案した KnowItNow や，Ghahramani ら[5]が提案した Bayesian Sets，Wang ら[6]の提案した SEAL がある．KnowItNow や SEAL は，Web ページをコーパスとして用いている．また，Google の実験的な試みとして行われている，Google Sets[4]も関連研究としてあげられる．

2.1. Tchai の概要

我々の手法は Tchai アルゴリズムに基づいているので，Tchai アルゴリズムの詳細について述べる．

Tchai はクエリログをコーパスとして意味知識獲得を行うアルゴリズムである．クエリログの中でも，バイワードクエリであるもののみを抽出対象としている．バイワードクエリとは，“東京 天気”のように，半角スペースで区切られた 2 つの文字列からなるクエリのことをいう．ユーザによっては複数の半角スペースや，全角スペースなどで区切ることもあるが，それらはすべて 1 つの半角スペースへ整形してコーパスに用いるようにする．また，Tchai におけるパターンとインスタンスは図 1 のように定義される．

クエリ	天気 東京
パターン	天気 #, # 東京
インスタンス	天気，東京

図 1. Tchai におけるパターンとインスタンス

これを見るとわかるように，1 つのバイワードクエリから，パターンとインスタンスがそれぞれ 2 つずつ抽出できる．

インスタンスは，パターンによって抽出される，あるカテゴリに属するキーワードとして定義される．このカテゴリはあらかじめ与えられているものとする．

パターンは，バイワードクエリにおける右または左をワイルドカード(“#”)とした，インスタンスの抽出パターンとして定義される．例えば，“天気 #”というパターンに対しては，“神奈川”，“千葉”，“埼玉”などのインスタンスがマッチする可能性がある．したがって，適切なパターンを選択することで，同じカテゴリに属するインスタンスが抽出できるというのが，Tchai アル

[1] http://www.google.co.jp/
[2] http://www.google.co.jp/maps
[3] http://www.yahoo.co.jp/
[4] http://labs.google.com/sets

ゴリズムの基本的な考え方となっている.

Tchai は次に示す 8 ステップでクエリログからインスタンスの抽出を行う.

1. シードインスタンスの入力
2. 表層パターン P の抽出
3. P の信頼度計算
4. 信頼度上位 k パターンの選択
5. インスタンス I の抽出
6. I の信頼度計算
7. 信頼度上位 m インスタンスの出力
8. k の値を 1 増やし，3 へ戻る

図 2 にこの Tchai の流れを概要図として示す．これを見て分かるとおり，表層パターンの再抽出は行われず，そのかわりにシードから抽出されたパターンの信頼度を再計算するようになっている.この変更により，Espresso において問題であった実行速度の遅さを Tchai は改善している.

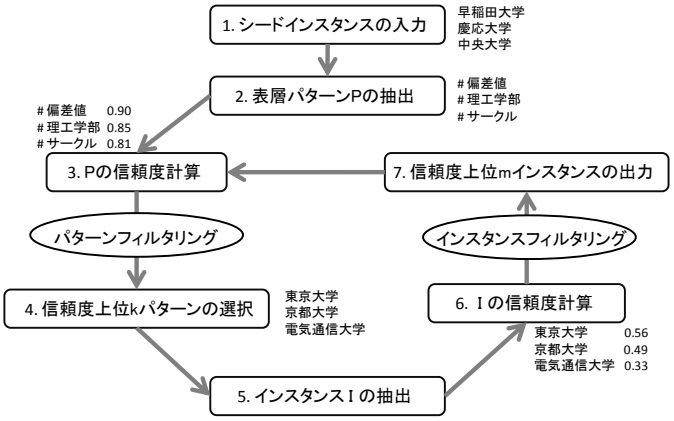

図 2. Tchai の概要図

インスタンス i の信頼度 $r_i(i)$ およびパターン p の信頼度 $r_\pi(p)$ は次の計算式で表現される.

$$r_i(i) = \frac{\sum_{p \in P} \frac{pmi(i,p)}{local\ max\ pmi} r_\pi(p)}{|P|}$$

$$r_\pi(p) = \frac{\sum_{i \in I} \frac{pmi(i,p)}{local\ max\ pmi} r_i(i)}{|I|}$$

$r_i(i)$ および $r_\pi(p)$ は再帰的に定義されており，インスタンスならパターンの，パターンならインスタンスの前回の信頼度を利用して信頼度を再計算していることがわかる．なお，初期値はどちらも 1 で定義される.

また，どちらの信頼度においても PMI という共起の強さを測る指標が用いられている．この式は今回のタスクでは次のように表現される.

$$pmi(i,p) = log \frac{N|i,p|}{|i,*||*,p|}$$

ここでのアスタリスクはワイルドカードを示し，どんなインスタンス(パターン)も入りうることを表している.また, N はバイワードクエリ総数を示している.このように, Tchai はパターンとインスタンスにおける共起の強さを計算することで, 信頼度を決定している.

3. 提案手法

Tchai のような, ブートストラッピングを用いた情報抽出手法には，意味ドリフト(Semantic Drift)という問題がつきまとう．意味ドリフトとは，インスタンスを抽出するためのパターンを選択するさいに，誤って共起インスタンス数の多いパターン(ジェネリックパターン)を選択することで, 抽出されるインスタンスが本来のカテゴリからずれてしまう現象のことをいう．意味ドリフトの例を図 3 に示す.

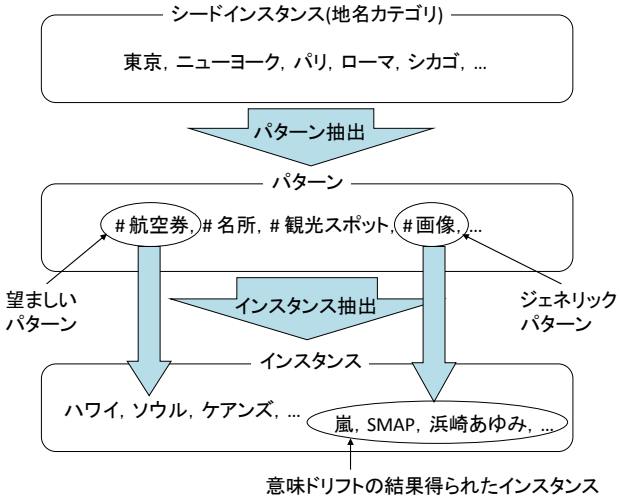

図 3. 意味ドリフトの例

Tchai は，既に抽出したパターンのうち，最も共起インスタンス数が多いパターンの共起インスタンス数の 2 倍以上のインスタンスを獲得するパターンや，既に抽出したインスタンスのうち，最も共起パターン数が多いインスタンスの共起パターン数の 2 倍以上のパターンを獲得するインスタンスを選択しないことで，この問題を回避する試みを行っている．しかしながら，論文ではこの閾値設定の理由について触れられていなかった.

そこで，我々は Tchai をベースとして，パターン選択方法の変更を行った．この変更により，インスタンス抽出精度の向上を目指した.

3.1. アプローチ

Tchai は共起数に着目して意味ドリフトを防ぐ試みを行っていたが,我々は“パターンを抽出するために利用したインスタンス”と “パターンによって抽出されるインスタンス”との関係に着目して意味ドリフトを防ぐ試みを行った．つまり，今までに抽出されたインスタンスと，これから抽出しようとしているインスタンスが，ある程度似ているときだけパターンを採用するという,パターンフィルタリングを行ったのである．このフィルタリングによって，既に抽出されたインスタンスと，抽出しようとしているインスタンスのカテゴリの相違が起こりにくいことが期待でき，意味ドリフトを防ぐことができると考えられる．

本研究では，フィルタリングにおいて 2 つの提案手法を提示する．なお，以下では既に獲得したインスタンス集合を X，パターン選択によって得られるインスタンス集合を Y としている．

3.2. Simpson 係数によるフィルタリング

Simpson 係数は, Jaccard 係数やコサイン距離と並び,共起の強さを測る尺度としてよく用いられる．今回 Simpson 係数を選択したのは，X と Y の大きさを比較したときに，X の大きさの方がほとんどの場合大きくなるからである．Simpson 係数は分母に$|X|$と$|Y|$の最小値をとるので，反復を繰り返して X が大きくなっても指標の値を一定に保つことができる．この Simpson 係数を用いた手法を，以降，提案手法①と呼ぶことにする．

なお，今回用いた Simpson 係数は共起頻度でフィルタリングをかけたものを用いている.以下に式を示す.

$$Simpson(|X|,|Y|)=\frac{|X\cap Y|}{Min(|X|,|Y|)}$$
$$\text{if}\quad |X\cap Y|<n \quad \text{then}\quad 0$$

3.3. パターン安定度 SOP (Stability Of Pattern)によるフィルタリング

今回のタスクにおいては，可能な限りインスタンスを抽出することが求められており，X と Y の共起が高ければ良いわけではない．X と Y の共起が高いだけでは，既に抽出したインスタンスとほとんど同じような集合を選択することになり，新たなインスタンスを獲得できる可能性が少なくなってしまう．したがって，意味ドリフトが起きない範囲で，なるべく既出でないインスタンスが獲得できるような集合を選択する必要がある．

そこで，X と Y の重なりが調度半分となるときに最も値が高くなるような指標，SOP (Stability Of Pattern)を用いることを考えた．この指標により，意味ドリフトを防ぎつつ新たなインスタンスも数多く獲得されることが期待できる．この SOP を用いた手法を，以降,提案手法②と呼ぶことにする．なお，SOP は次のように定式化される．

$$SOP(|X|,|Y|)=\frac{|X\cap Y|}{|Y|}\cdot\frac{|Y\setminus X|}{|Y|}$$
$$or\quad SOP(|X|,|Y|)=\frac{|X\cap Y|}{|X|}\cdot\frac{|X\setminus Y|}{|X|}$$
$$\text{if}\quad |X\cap Y|<n \quad \text{then}\quad 0$$

4. 評価実験

4.1. 実験設定

評価実験は次のような設定のもとで行った．

- コーパス：日本語向け Web 検索エンジン 2008 年 6 月から 2009 年 6 月まで約一年分のクエリログのうち，バイワードクエリの頻度上位 1 万件
- 正解データ:日本語向けモバイル Web 検索エンジンクエリログ 2008 年度分のうち，頻度上位 3 万件を人手でカテゴリ分類したもの
- カテゴリ：企業，健康・医療，番組名，芸能人の 4 カテゴリ

実行にあたっては反復を 10 回繰り返し，反復毎に最大 200 インスタンスを獲得するようにした.そして,正解データを用いて 4 カテゴリでの実行結果の適合率(Precision)，再現率(Recall)，F 値(F-measure)の平均値を算出し，既存手法 Tchai と比較を行った．適合率，再現率，F 値はそれぞれ次のような式で求めている．

$$Precision=\frac{R}{M}$$
$$Recall=\frac{R}{C}$$
$$F-measure=\frac{2\cdot Precision\cdot Recall}{Precision+Recall}$$

M は抽出されたインスタンスの総数であり，R はそのうち正解であったものの数である．また，C は正解カテゴリに含まれるインスタンスの総数である．F 値は上式の通り，適合率と再現率の調和平均で定義される．なお，Simpson 係数は 0.5，SOP は 0.23，共起頻度 n は 10 でフィルタリングを行った．これらの数値を高い値に設定するとフィルタリングを通過できないパターンが増えてしまう．また，低い値に設定するとフィルタリングの意味をなさなくなる．今回の実験では，予備実験の結果，それぞれの手法で良好な結果を得た上記の値を利用した．

4.2. 結果と考察

4.2.1. カテゴリごとの結果と考察

企業カテゴリ(図 4)では，Tchai に比べて提案手法が適合率でどちらも大きな値を示している．企業カテゴリにおけるシードは，{楽天，アマゾン，ANA，JAL，NHK}であった．ANA や JAL は企業名ではあるが，旅行のイメージも大きい．そのため，Tchai では旅行カテゴリへの意味ドリフトが発生し，適合率を大きく下げる結果となった．

健康・医療カテゴリ(図 5)では，提案手法が優位であることが分かるものの目立って大きな差とはならなかった．これは，健康・医療カテゴリが，ある意味で閉じたカテゴリとなっており，他のカテゴリへ意味ドリフトが起きにくいからだと考えられる．

番組名カテゴリ(図 6)では，提案手法②が適合率は下げたものの，再現率で高い値を示した．これは，提案手法②が正解インスタンスも不正解インスタンスも数多く取得したことを示している．正解インスタンスを多く獲得したことで再現率を上げ，不正解インスタンスを多く獲得したことで適合率を下げたと考えられる．SOP では，意味ドリフトを防ぐとともに，数多くのインスタンスを獲得しようとするような指標となっているため，このような結果になったと考えられる．

芸能人カテゴリ(図 7)も，健康・医療カテゴリと同様に，提案手法の優位性は示されているものの，それほど大きな差は見られなかった．これは，芸能人カテゴリのインスタンスが，画像や動画，ブログなどのジェネリックパターンと非常に共起しやすいためである．つまり，他のカテゴリからすればジェネリックパターンであるこれらのパターンが，芸能人カテゴリにとってはジェネリックパターンとはならないのである．したがって，パターンフィルタリングがそれほど意味をなさなくなり，このような結果となったと考えられる．

4 カテゴリ平均での結果と考察

表 1 と図 8 を見ると分かるとおり，提案手法①が最も高い適合率を示す手法となった．既存手法の Tchai よりも約 1.17 倍の精度向上をあげることができている．一方，再現率においては Tchai より約 1.19 倍高いものの，提案手法②に比べると低い値となってしまった．

これは，稀なパターンが選択されてしまった場合に指標値が高くなるという Simpson 係数の特徴が影響していると考えられる．稀なパターンを選択したことで Y が小さくなると，Simpson 係数における分母が小さくなってしまい，共起しているものの数が少ないにも関わらず，指標値は大きくなってしまうのである．結果として，反復するごとに，選択されるパターンが

		Tchai	提案手法① Simpson	提案手法② SOP
企業	適合率(%)	17.19	46.60	35.36
	再現率(%)	12.84	13.36	19.56
	F値(%)	14.70	20.77	25.19
健康・医療	適合率(%)	63.53	67.61	66.76
	再現率(%)	36.10	39.75	49.07
	F値(%)	46.04	50.07	56.56
番組名	適合率(%)	41.90	41.77	36.52
	再現率(%)	34.65	50.00	63.64
	F値(%)	37.93	45.52	46.41
芸能人	適合率(%)	79.15	79.15	82.78
	再現率(%)	21.62	21.99	25.59
	F値(%)	33.96	34.42	39.09
合計	適合率(%)	50.44	**58.78**	55.36
	再現率(%)	26.30	31.28	**39.46**
	F値(%)	33.16	37.69	**41.81**
	適合率(比)	1.00	**1.17**	1.10
	再現率(比)	1.00	1.19	**1.50**
	F値(比)	1.00	1.14	**1.26**

表 1. 提案手法と既存手法の比較

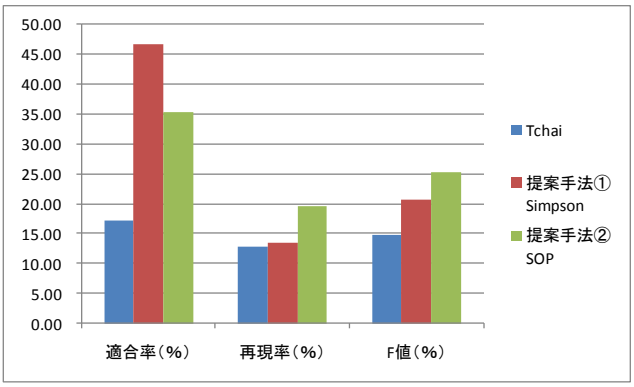

図 4. 企業カテゴリの結果

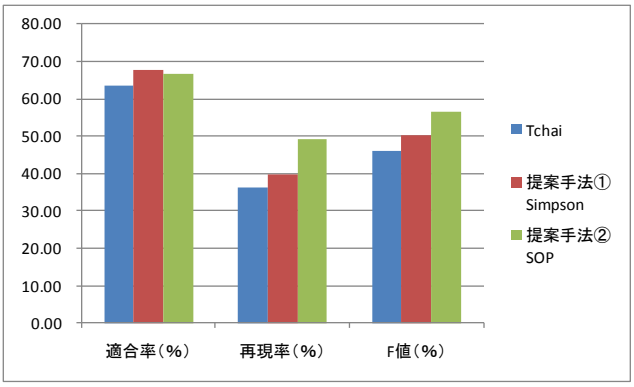

図 5. 健康・医療カテゴリの結果

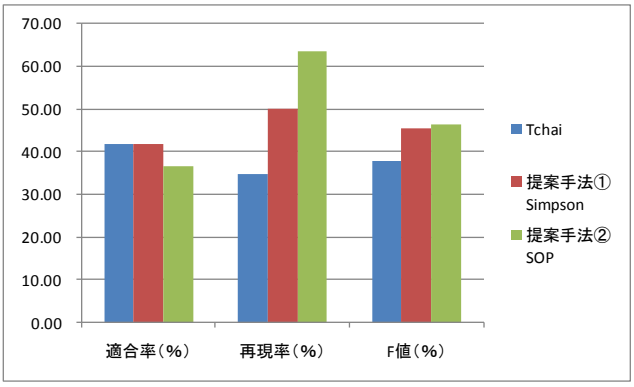

図 6. 番組名カテゴリの結果

しだいに稀なものへと遷移してしまい，抽出できたインスタンスの総数が減少してしまう．これにより，正解インスタンスを網羅的に取得できず，再現率が低下したのである．またその一方で，共起の強いパターンのみを選択し続けることで不正解インスタンスの抽出が抑えられ，適合率は向上したと考えられる．

提案手法②は，再現率，F 値が最も高い手法となった．ジェネリックパターンが選択されると，$|X \cap Y|$が大きい値となるのでフィルタされる．逆に，稀なパターンが選択されると，$|X \cap Y|$が小さい値となるのでフィルタされる．この SOP の適切なフィルタによって，ジェネリックでも稀でもないパターンがうまく選択された．その結果，意味ドリフトを抑えつつ多くのインスタンスが獲得できたため，このような結果になったと考えられる．

5. おわりに

本研究では，クエリログをコーパスとした，ブートストラッピングによる意味知識獲得の改善手法を提案した．パターン選択時にフィルタリングをかけることによって，意味ドリフトを抑えつつ，多くの正解インスタンスを抽出することができた．

我々は，Simpson 係数を用いてフィルタリングをかける手法と，SOP によってフィルタリングをかける手法を提案し，既存手法との比較実験を行った．

Simpson 係数を用いた手法では，ベースシステム Tchai よりも適合率で約 1.17 倍の精度向上をあげることができた．不正解インスタンスの混入を避けたいタスクにおいては，この手法が有効であると考えられる．一方，再現率は SOP を用いた手法より少し劣るので，正解インスタンスを数多く取得したいタスクには向かないと考えられる．

SOP を用いた手法では，再現率，F 値において最も良い成果をあげることができた．特に，F 値では Tchai より 1.26 倍高い精度での抽出を行うことができた．

SOP を適用することで，ブートストラッピングにおいて特有の意味ドリフト問題と，共起頻度を測る指標において特有の稀な共起を過大評価する問題を，うまく回避したことが精度向上につながった．今回は Tchai における適用であったが，ブートストラッピングを用いる他の手法においても SOP が有効であるのではないかと考えられる．

今後の課題としては，バイワード以外のクエリへの適用があげられる．また，インスタンスひとつに対して複数のカテゴリを付与するといったことも課題としてあげられる．

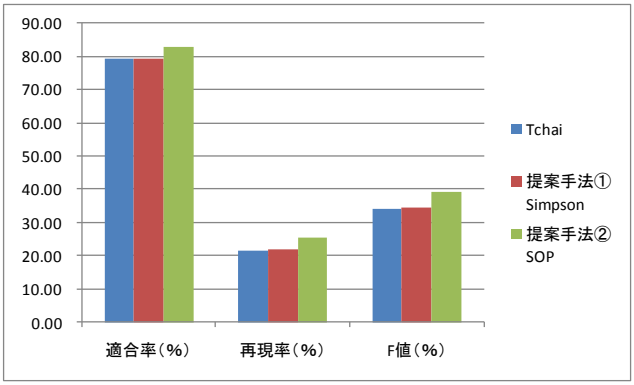

図 7. 芸能人カテゴリの結果

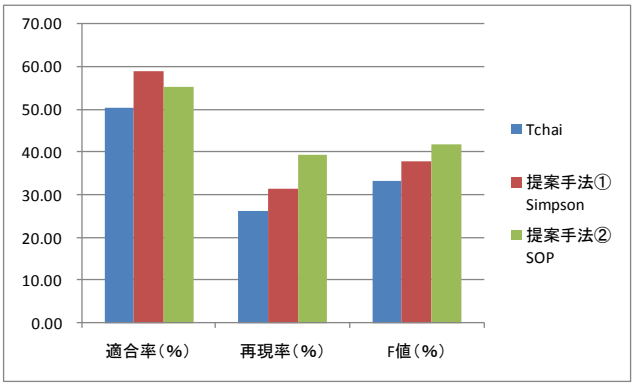

図 8. 提案手法と既存手法の比較

文　献

[1] Patrick Pantel, Marco Pennacchiotti, “Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations”, Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, pp. 113-120, 2006

[2] 小町守，鈴木久美，“検索ログからの半教師あり意味知識獲得の改善(Improving Semi-supervised Acquisition of Semantic Knowledge from Query Logs)”，人工知能学会論文誌，23 巻 3 号，pp. 217-225, 2008

[3] M. J. Cafarella, D. Downey, S. Soderland, and O.Etzioni, “KnowItNow: Fast, Scalable Information Extraction from the Web”, in EMNLP, 2005.

[4] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, “Unsupervised Named-Entity Extraction from the Web: An Experimental Study”, Artificial Intelligence, vol. 165, pp. 91-134, 2005.

[5] Z. Ghahramani and K. A. Heller, “Bayesian Sets”, in Advances in Neural Information Processing Systems, 2005.

[6] Richard C. Wang and William W. Cohen, “Language -Independent Set Expansion of Named Entities using the Web”, In Proceedings of IEEE International Conference on Data Mining (ICDM 2007), Omaha, NE, USA. 2007.