Vol.2009-NL-191 No.9
Vol.2009-SLP-76 No.9
2009/5/21

ウェブ検索ログを用いたラベル伝播による意味カテゴリ獲得

小町　守[†1]　牧本慎平[†2]
内海慶[†2]　颯々野学[†2]

近年ウェブ検索が一般的になり，ウェブを用いた知識獲得の研究が盛んになってきている．検索ログはユーザのユーザの関心を反映した情報源であり，ターゲット広告や検索支援のための情報抽出源として注目を集めている．しかしながら，既存の検索クエリログを用いた意味カテゴリ学習の研究ではユーザが入力したクエリを用いることによるリソースの問題，ブートストラップに代表される手法の可搬性の問題，そしてウェブを対象にした大規模データに対する拡張性の問題，の 3 つの問題点があった．そこで，本研究では検索クリックスルーログを用いた高精度な意味カテゴリ獲得，そしてラベル伝播によるグラフ理論に基づく手法の提案，最後に MapReduce を用いた並列分散計算により，これらの問題を解決する．意味カテゴリ学習タスクにおいて検索クリックスルーログを用いた研究はこれまでになく，本研究では既存手法に比べ高精度・高再現率で意味カテゴリを獲得できることを示した．

Learning Semantic Categories from Web Search Logs Using Label Propagation

Mamoru Komachi,[†1] Shimpei Makimoto,[†2] Kei Uchiumi[†2]
and Manabu Sassano [†2]

Recently knowledge acquisition from the web is actively studied as web search becomes widespead. Web search logs are getting much more attention than before for the purpose of information extraction for applications such as targeted advertisement and query suggestion. However, previous work has three problems: query logs of poor quality due to inherent variations of the users' input; portability issue resulted from machine learning method such as bootstrapping; and scalability issue for handling large amount of data extracted from the web. Thus, we propose to use web clickthrough logs for learning semantic categories at high precision; apply a graph theoretic label propagation method; and perform parallel and distributed computation using MapReduce. Our main contribution is the use of web clickthrough logs for the task of learning semantic categories, and we demonstrated it achieves higher precision and recall than previous work.

1. はじめに

近年ウェブ検索が一般的になり，ウェブを用いた知識獲得の研究が盛んになってきている．特に検索ログはユーザのユーザの関心を反映した情報源であり，ターゲット広告や検索支援のための知識獲得源として注目を集めている．

検索クエリは，一般的なテキストと比較して検索を行うユーザの関心を反映している[7] ため，検索に関するタスクにおいては，検索クエリログから学習した意味カテゴリが有効であると考えられる．そこで，自然言語処理において検索ログを活用したさまざまな知識獲得が試みられてきた．たとえば，小町ら[10] は大規模な日本語の検索クエリログを用い，Pantel らが提案した *Espresso*[5] に基づく検索クエリログを対象としたブートストラップによる意味カテゴリ学習方法 *Tchai* を提案した．

しかしながら，この先行研究には以下の 3 点の問題があった．

リソースの問題　先行研究では，検索クエリのログをそのまま使用するため，ユーザが知りたい情報を反映する単語の情報は取得できるものの，実際その単語が知りたいカテゴリを絞り込むものとして適切ではない場合があった．そのため，検索クエリログだけから精度高く認識することは困難であった．

可搬性の問題　関係抽出や固有表現抽出などの知識獲得タスクで用いられている *Espresso*[5] や，検索クエリログを対象とした *Tchai*[10] では，使用するデータに合わせて 8 個以上の変数を設定しなけ

†1 奈良先端科学技術大学院大学情報科学研究科
630-0192 奈良県生駒市高山町 8916-5
†2 ヤフー株式会社
107-6211 東京都港区赤坂 9-7-1 ミッドタウン・タワー

Vol.2009-NL-191 No.9
Vol.2009-SLP-76 No.9
2009/5/21

ればならなかったが，これらの変数の設定によって大きく性能が変化してしまうため，現実的な問題に適用する際には最適なパラメータの調整が必要であり，実運用の障害となっていた．

拡張性の問題 ブートストラップをはじめとする先行研究は 1 台の計算機で実行することが前提となっているアルゴリズムなので，大規模化することが困難であった．

そこで，本研究ではそれぞれの問題に対し，以下の 3 つの解決方法を提案する．

① **検索クリックスルーログ**の活用
② **ラベル伝播**による半教師あり学習
③ **並列分散処理**を用いた大規模化

まずリソースの問題について，先行研究のように検索クエリログを用いるだけではなく，検索クリックスルーログを用いた意味カテゴリ学習を提案する．検索クエリを入れて検索エンジンが返す結果に対し，ユーザがタイトル・アドレス・要約（スニペット）を見てクリックしたアドレス（**検索クリックスルー**）は，ユーザの意図を直接表していると考えられる．つまり，同じアドレスに到達する検索クエリは同じ意図で検索された可能性が高いと考えられる．特に同一のページに到達する異なる 2 つの検索クエリは同義語であることが多く，クリックスルーログを用いることによって，カテゴリの学習を高精度に行えることが予想される．提案手法では検索クリックスルーログに加えて検索クエリログもパターンとして用い，精度と再現率両方を向上させることを目的としている．

次に可搬性の問題について，提案手法はラベル伝播を用い，少ないパラメータ数で従来手法と同程度の性能を達成する．特に，検索クエリとパターンをそれぞれノードとし，エッジには検索クエリとパターンの共起スコアが付与されているような 2 部グラフを考えると，*Espresso* や *Tchai* といったブートストラップ手法はこの 2 部グラフ上の関連度の計算と見なすことができる．[1] 提案手法もこうしたアルゴリズムの一種であるが，ブートストラップに比べ調整するべきパラメータ数が少ないという利点がある．

最後に拡張性の問題について，本研究では MapReduce を用いて並列分散処理を行うことで解決する．グラフ理論に基づくアルゴリズムには並列処理が可能なアルゴリズムが多数存在し，それらを用いることによって効率的に計算できる．

我々の手法は検索ログからの意味カテゴリ学習を少数のラベル付きシードを与えるラベル伝播によって学習する問題として扱う．意味カテゴリ獲得のために日本語の検索ログから半教師あり学習を行う手法としては，ブートストラップアルゴリズムを用いた[10] があるが，自然言語処理において検索クリックスルーログを活用した研究は我々の知る限りこれが初めてのものである．

2. 関 連 研 究

近年検索ログを用いた知識獲得が盛んになってきた．これらのアルゴリズムの特徴は，あるカテゴリや関係を抽出するための文脈パターンを用い，そのカテゴリに属する抽出対象の**インスタンス**（例：動物クラスの**ネコ**），もしくは属性抽出の場合は特定の関係にある単語のペア（例：**会社**に対する**社長**）を学習することである．本研究では特定の意味カテゴリに属する固有表現の抽出を目標としている．

自然言語処理分野における検索ログの利用は Paşca et al. が先鞭をつけ，特に検索クエリログからの固有表現に関する知識獲得の手法を提案している．[2),3)] 彼らは固有表現の属性を学習することに焦点を当てているので，本研究とは目的が異なっている．

また，検索クエリログに加えてウェブ文書を用いて意味カテゴリ学習を行う手法の研究もなされている．[4),8)] Talukdar et al. による研究[8] は少数のシードを与えてラベル伝播により意味カテゴリを学習する点において共通しているが，我々の研究は検索クエリログに加えて検索クリックスルーログを使う点とラベル伝播を行う 2 部グラフの作成方法が異なる．

萩原ら[11] は日本語の検索クエリログを用いたグラフカーネルに基づく意味的類似度計算をクエリ書き換えのタスクに適用した．彼らの手法とはグラフに基づいているという点は共通しているが，我々は検索クリックスルーログを使っているという点と，類似度行列の作成方法，および意味カテゴリ学習のタスクに用いたという目的が異なる．

小町らの提案する *Tchai*[10] は検索クエリログを用いたブートストラップ手法の一種で，特定のカテゴリに属する少数のシードインスタンスからスタートし，以下の式によってインスタンス獲得とパターン抽出を交互に繰り返しつつ，大規模な意味カテゴリを学習する．彼らの手法は検索クエリログに特化して高い精度で意味カテゴリ学習を行うことができるが，ブートストラップ手法はパラメータ数が多いため調整が難しく，大規模化が困難であるという問題があるほか，検索クリックスルーを使用せず，検索クエリログのみを対象にしている点が異なる．

3. *Quetchup*⋆1 アルゴリズム

本節ではラベル伝播に基づく検索クエリの意味カテゴリ学習手法について説明する．我々はこのアルゴリズムを *Quetchup*（ケチャップ）と名付けた．

3.1 ラベル伝播による半教師あり学習

本手法が用いるラベル伝播は 9) に基づいている．ラベル伝播をはじめとするグラフに基づく半教師あり手法は，少数のシードを用いても比較的高い精度が得ら

⋆1 Query Term Chunk Processor

Vol.2009-NL-191 No.9
Vol.2009-SLP-76 No.9
2009/5/21

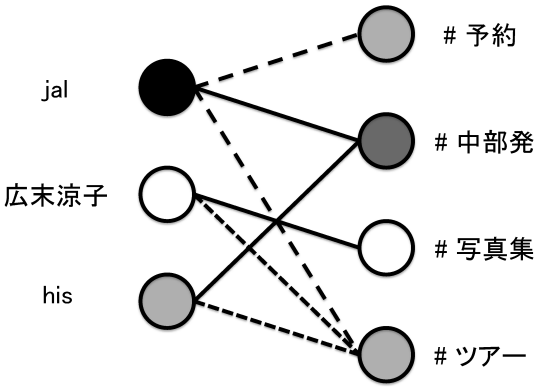

図 1 シード単語からラベルが伝播していく

入力：
シードインスタンスベクトル $F(0)$
インスタンス類似度行列 A
出力：
インスタンススコアベクトル $F(t)$
1: A を $D^{-1/2}(A)AD^{-1/2}(A)$ により正則化する
2: 正則化ラプラシアン行列 $L = D(A) - A$ を作成する
3: 収束するまで $F(t+1) = \alpha LF(t) + (1-\alpha)F(0)$ を繰り返す

図 2 正則化ラプラシアンを用いたラベル伝播

れ，また大規模化が容易であるという特徴がある．行列の固有値計算を伴う手法はインスタンス数を n とするとナイーブな実装では $O(n^3)$ の計算量がかかるが，近似により $O(n^2t)$ の計算量（ただし t はステップ数）で解を求めることができる．

図 1 は旅行ドメインにおいてシード単語「jal」を用いたときのラベル伝播の様子を描いたものである．グラフは左側のノードが単語，右側のノードがその単語と共起するパターンとなっている 2 部グラフで，線の強さが共起の度合いを示し，ノードの濃さが旅行ドメインらしさを表している．「jal」と共起の度合いが強いパターン「♯ 中部発」[⋆1]は旅行ドメインに特徴的なパターンであるとしてラベルが伝播する．一方，「♯ ツアー」というパターンは他の「広末涼子」という旅行ドメインではない単語とも共起するため，比較的中立なパターンである．単語「his」は「♯ 中部発」と「♯ ツアー」という 2 つのパターンを「jal」と共有しているため，シード単語「jal」のラベルが伝播して旅行ドメインらしい単語と分類される．

このようにラベル伝播ではシードとして与えるノードのラベルを順次隣接ノードに伝播していく手法であり，最適なラベルはラベル伝播のプロセスが収束した状態におけるラベルとして与えられる．

⋆1 ♯ は単語の入る位置を示す．

正則化ラプラシアンを用いたラベル伝播手法を図 2 に示す．ただし，**インスタンス類似度行列** A は**インスタンス・パターン行列** W を用いて $A = W^TW$ とする．W_{ij} はインスタンス x_i とパターン p_j の共起回数を用い，$D^{-1}(W)W$ によって正規化する．また，$D(N)$ は $D(N)_{ii} = \sum_j N_{ij}$ で定まる行列 N の次数対角行列である．

事例 $x_i \in \mathcal{X}$ に対し $F(0)$ はラベル集合 $y_i \in \mathcal{Y}$ から作成される x のスコアベクトルであり，$F(t)$ は t ステップ終了時の x のスコアベクトルである．1 カテゴリの学習をする場合，シードとして与えるインスタンスに対応する値を 1 とし，ラベルが不明なそれ以外のインスタンスに対応する値を 0 とする．この場合，出力として得られるインスタンスのスコアベクトルは，シードとして与えたインスタンスに対する類似度順に整列したベクトルになっている．また，2 カテゴリの学習をする場合，シードとして与えるインスタンスはそれぞれ 1 または -1 の値を与え，最終的なスコア y_i の符号の正負によって x_i のラベルを決定する．[⋆2]

ラベル伝播手法はシードのラベルとグラフ構造どちらを重視するかというパラメータ $\alpha \in [0, 1)$ を持ち，α が 0 に近づけばシードのラベルに偏った結果となり，α が 1 に近づけばラベルなしデータから作成されるグラフ構造を考慮した結果となる．

3.2 ラベル伝播計算の効率化

ラベル伝播は一般的には類似度行列 A に基づき計算するものであるが，ウェブデータを対象とした知識獲得では，インスタンス数が 100 万から 1,000 万のオーダーになることも珍しくない．そのため，大規模データに対しては単純に類似度行列を作成すると計算量の面 [⋆3]からも記憶領域の面からも非現実的なので，いくつか計算上の工夫が必要である．

3.2.1 記憶領域の削減

まず記憶領域について述べる．類似度行列は $O(n^2)$ の記憶領域が必要であり，大規模なデータを対象にした類似度行列の保持は非現実的である．[⋆4]そこで，類似度行列 A を $A = W^TW$ として 2 つの行列に分解して保持し，毎回計算することで，記憶領域の爆発を抑える．A は密行列なのに対し，インスタンス・パターン行列 W は疎行列なので，このとき必要な記憶領域は 1 インスタンス当たりの平均共起パターン数を p とすると $O(np)$ であり，$n \gg p$ なので，現実的な

⋆2 3 つ以上の n 個のカテゴリを学習する場合はシードとしてベクトルではなく n 次元の行列を作成し，各事例 x_i はラベル $y_i = \arg\max_j F(t)_{ij}$ でラベルづけする．

⋆3 数 GB 程度のメモリを搭載したワークステーションで実行できる固有値計算は，アイテム数がせいぜい数万程度までであり，行列の分解やクラスタリングによる次元縮約を行わないと大規模データは扱えない．

⋆4 たとえば類似度を 4 ビットで保持したとしても 16GB のメモリで計算できる類似度行列のサイズはせいぜい 10 万程度である．

Vol.2009-NL-191 No.9
Vol.2009-SLP-76 No.9
2009/5/21

サイズに落とし込むことができる．[⋆1]

3.2.2 近似による計算量の削減

次に計算量について述べる．ラベル伝播計算は

$$F^* = \sum_{t=0}^{\infty}(I+\alpha L)^t = (I-\alpha L)^{-1}$$

によって求められ，この計算は固有値計算を伴うため計算量 $O(n^3)$ かかるが，ラベル伝播のステップ数を定数回に止めることで F^* を近似することができる．ラベル伝播のステップ数 t は現在のノードから何歩先のノードまで情報を伝播させるかに対応し，t を増加させるとグラフ上のあらゆるパスを考慮に入れるが，t を 0 に近づけると現在のノードの近傍のみを考慮に入れることに対応する．

このようにラベル伝播を近似すると MapReduce を用いた並列分散計算を行うことができるという利点がある．

4. ウェブ検索ログを用いた実験

本節では先行研究[10] と提案手法 *Quetchup* の比較実験について述べる．

4.1 実験設定

検索ログ インスタンス獲得に用いた知識獲得源は，Yahoo! 検索で 2008 年 8 月に検索された検索ログ集合のうち，異なり頻度上位 1,000 万クエリを用いた．[⋆2]検索クエリには検索ログ中に現れた総頻度が振られている．

インスタンス-パターン行列の作成 クリックスルーパターンとして，クエリに対してクリックされたアドレスをパターンとして用いた．クリックされた回数が 200 回以下のアドレスは削除し，また，共起するクエリが 1 つだけのアドレスも削除した．[⋆3]

クエリパターンとしては，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を用いた．たとえば単語「jr」に対して 2 単語クエリ「jr 時刻表」から文脈パターン「♯時刻表」を得る．共起する頻度が 100 回以下の文脈パターンは削除した．インスタンス・パターン行列 W の要素 W_{ip} は $W_{ip} = \mathrm{count}(i,p)$ で与えられる．

対象カテゴリ 今回のタスクは 10) にならい，「旅行」カテゴリと「金融」カテゴリの 2 カテゴリを対象とした．ウェブ検索ユーザの関心はタスク依存なので，彼らと同じく少数のシード単語から単語カテゴリを学習することにした．[⋆4]

表 1 各カテゴリでのシード単語

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

クエリに表記揺れや綴り誤りがある場合，代表表記に復元したうえで意味カテゴリを付与した．また，1 クエリに複数単語が含まれる場合，いずれかの単語が対象となる意味カテゴリに属するのであれば，全体をその意味カテゴリと判定した．[⋆5]

システム *Tchai* と *Quetchup* には同じシード単語を与えた．用いたシード単語は表 1 に示した．*Tchai* は 10) と同じパラメータを使用し，1 回当たり 10 インスタンスずつ取得し，10 回反復を繰り返して合計 100 インスタンスを獲得した．*Quetchup* の反復回数は 10 回，パラメータ α は特に断りがないかぎり 0.0001 とした．MapReduce の実装としてはオープンソースで開発されている Hadoop [⋆6] を用いた．

評価 意味カテゴリ学習タスクでは真の正解を定めることが困難なので，システムの評価には**順位 k での精度**（precision at k）と**相対再現率**[6] を用いた．[⋆7]相対再現率とは，あるシステムの出力を他のシステムがどれくらいカバーできるかを調べたものであり，次式で与えられる．

$$R_{A|B} = \frac{R_A}{R_B} = \frac{C_A/C}{C_B/C} = \frac{C_A}{C_B} = \frac{P_A \times |A|}{P_B \times |B|}$$

ここで $R_{A|B}$ はシステム B を基準としたシステム A の相対再現率であり，C_A, C_B はそれぞれシステム A, B が出力した正解の個数，C は真の正解の個数である．C は分子と分母で共通しているため相殺することにより，システム A, B の精度 P_A, P_B とシステムが出力したインスタンスの個数 $|A|, |B|$ により相対再現率を求めることができる．

4.2 実験結果

4.2.1 検索クリックスルーログの有効性

図 3 から図 6 までは検索クリックスルーログが精度と再現率向上に有効であることを示すため，検索クエリログと検索クリックスルーログの 2 つを用いた提案手法および *Tchai* により順位 k 番目での精度と相対再現率を描いたものである．相対再現率はクエリログのみを用いたシステムをベースラインとして，それぞれクリックスルーログを用いたシステムと *Tchai* と

⋆1 random projection などの次元縮約手法を用いることによってさらに圧縮することができる．
⋆2 ただし，以下特に断りがないかぎり実験時間短縮のため頻度上位 100 万検索ログを用いて実験した．
⋆3 計算時間の短縮と行列のサイズの圧縮のために有効である．
⋆4 ただし，特に断りがないかぎり旅行カテゴリのみで比較実験を行った．
⋆5 実際の検索システムで用いる場合，綴り誤りが多数含まれるため，現実的な設定に近づけて評価した．
⋆6 http://hadoop.apache.org/
⋆7 獲得された単語のうち上位にあるものから順に人手で精査するという用途が典型的に想定されるため，順位 k 番目での精度が実用上重要である．

Vol.2009-NL-191 No.9
Vol.2009-SLP-76 No.9
2009/5/21

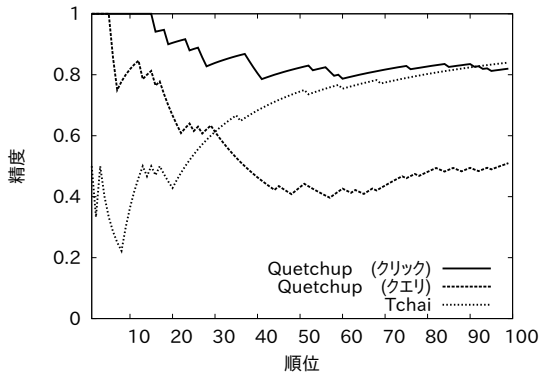

図 3 旅行ドメイン:*Quetchup* と *Tchai* の精度比較

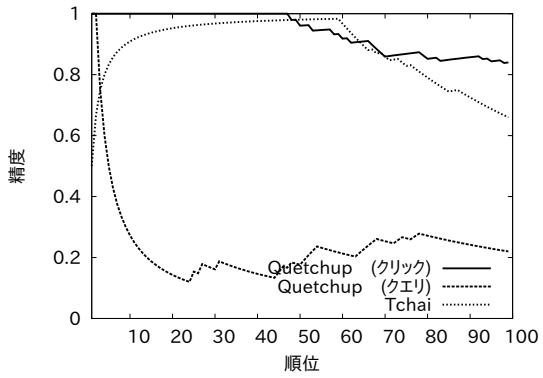

図 4 金融ドメイン:*Quetchup* と *Tchai* の精度比較

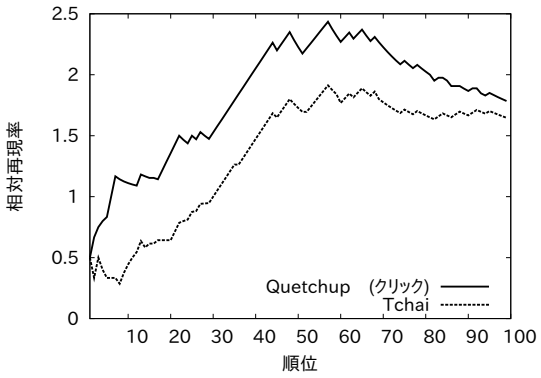

図 5 旅行ドメイン:*Quetchup* と *Tchai* の相対再現率比較

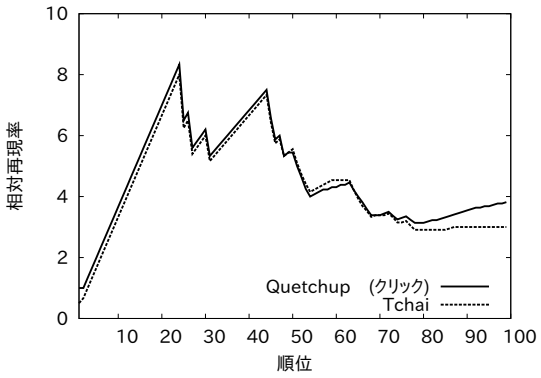

図 6 金融ドメイン:*Quetchup* と *Tchai* の相対再現率比較

を比較した．

検索クリックスルーを用いると精度と相対再現率が上昇することが確認され，検索クリックスルーログが意味カテゴリ学習タスクにおいて有効であることが分かった．クリックスルーログのみを用いた場合がもっとも精度が高く，上位 100 件を獲得しても精度の大幅な低下は見られなかった．また，相対再現率はクエリのみを用いたシステムに比べ 100 %を上回りであり，クリックスルーログによって高い再現率が得られることも分かった．

一方，クエリログのみを用いた場合は下位に行くに従って精度が下がった．実際に獲得されるクエリを分析したところ，アダルトクエリがもっとも多く，次にグルメに関するクエリ，就職に関するクエリ，住宅に関するクエリが続き，意味ドリフトが起きていることが確認された．これはインスタンスとパターンの共起スコアとして単純な頻度を用いているので，検索クエリとして頻度の高いクエリに高いスコアが割り振られたためだと考えられる．自己相互情報量や対数尤度比といった相対頻度を用いることにより，高頻度クエリの影響を抑えることができると考えられる．

また，旅行ドメイン・金融ドメインいずれにおいてもクリックスルーログを用いた提案手法は *Tchai* の精度を上回り，しかも順位が下位になっても精度の変化がほとんどないという利点がある．クエリログを用いた提案手法をベースラインとした場合の相対再現率を見ても提案手法はコンスタントに *Tchai* を上回っており，精度の面においても再現率の面においても提案手法は既存手法より優れていることが分かる．

4.2.2 獲得されたインスタンスとパターン

表 2 は検索クエリと検索クリックスルーそれぞれの特徴を示すため，獲得されるインスタンスとパターンのスコア上位 10 件を挙げたものである．

クリックスルーを用いた場合，異表記（じゃらん, ジャラン）や綴り誤り（jarann, jaran, じゅらん）であってもユーザは同じアドレス（http://www.jalan.net/）に到達できるので [⋆1]，これらのクエリのスコアが高くなっており，シード単語の同義語としてこれらのクエリを獲得することに成功している．

一方，クエリのみを用いてラベル伝播した場合，同義語以外のクエリも獲得できるようになるが，対象とする意味カテゴリに属さない単語（ad-box, アダコミ）も上位にランクインし，精度・再現率ともに下がってしまう．

⋆1 検索エンジン自体がこれらの表記揺れや綴り誤りに対し頑健で，適切なページを上位に表示しているためでもある．

表 2 獲得されたインスタンスとパターンのスコア上位 10 件

システム	インスタンス	パターン（アドレスから http:// は取り除いた）
Quetchup（クリックのみ）	じゃらん 宿泊, じゃラン, ジャラン, jarann, jaran, じゃらん net, jalan, じゅらん, ana 予約, ana.co.jp	www.jalan.net/, www.ana.co.jp/, www.his-j.com/ www.jreast.co.jp/, www.jtb.co.jp/, www.jtb.co.jp/ace/, www.westjr.co.jp/, www.jtb.co.jp/kaigai/, nippon.his.co.jp/, www.jr.cyberstation.ne.jp/
Quetchup（クエリのみ）	中部発, his 関西, 伊平屋島, ホテルコンチネンタル横浜, げんじいの森, フジサファリパーク, ad-box, アダコミ, スカイチーム, ノースウェスト	♯ 時刻表, ♯ 国内旅行, ♯ 宿泊, ♯ 北海道, ♯ 関西, ♯ 九州, ♯ マイレージ, ♯ 名古屋, ♯ 沖縄, ♯ 温泉
Tchai	静鉄バス, 相鉄バス, 函館バス, 大阪地下鉄, 琴電, 地下鉄御堂筋線, 芸陽バス, 新京成バス, jr 阪和線, 常磐線	♯ 時刻表, ♯ 路線図, ♯ 運賃, ♯ 料金, ♯ 定期, ♯ 運行状況, ♯ 路線, ♯ 定期代, ♯ 定期券, ♯ 時刻

5. ま と め

本研究ではラベル伝播を用いた検索ログからの意味カテゴリ学習手法 *Quetchup* を提案した．この手法の主要な貢献は，意味カテゴリ学習タスクにおける検索クリックスルーログの有用性を指摘したことと，ラベル伝播を用いたスケーラブルな意味カテゴリ学習法を提案したことである．提案手法は単語獲得の精度において既存手法より優れているだけでなく，ブートストラップに比べてパラメータが少ないため扱いが容易であり，並列分散計算環境を用いることによって大規模化することが可能である．

今後は 3 つ以上のドメインに対して本手法を適用し，性能を比較してみたい．また，検索クリックスルーログ以外でも自然言語処理に有用なログを活用し，高精度化とカバー率の向上に取り組んでいく予定である．

謝　　辞

本研究は日本学術振興会特別研究員奨励費の助成を受けたものである．

参 考 文 献

1) Mamoru Komachi, Taku Kudo, Masashi Shimbo, and Yuji Matsumoto. Graph-based Analysis of Semantic Drift in Espresso-like Bootstrapping Algorithms. In *Proc. of EMNLP-2008*, pp. 1010–1019, 2008.
2) Marius Paşca. Organizing and Searching the World Wide Web of Fact — Step Two: Harnessing the Wisdom of the Crowds. In *Proc. of WWW-07*, pp. 101–110, 2007.
3) Marius Paşca and Benjamin Van Durme. What You Seek is What You Get: Extraction of Class Attributes from Query Logs. In *Proc. of IJCAI-07*, pp. 2832–2837, 2007.
4) Marius Paşca and Benjamin Van Durme. Weakly-Supervised Acquisition of Open-Domain Classes and Class Attributes from Web Documents and Query Logs. In *Proc. of ACL-2008*, pp. 19–27, 2008.
5) Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In *Proc. of COLING-ACL*, pp. 113–120, 2006.
6) Patrick Pantel and Deepak Ravichandran. Automatically Labeling Semantic Classes. In *Proc. of HLT/NAACL-04*, pp. 321–328, 2004.
7) Craig Silverstein, Monika Henzinger, Hannes Marais, and Michael Moricz. *Analysis of a Very Large AltaVista Query Log.* Digital SRC Technical Note 1998-014, 1998.
8) Partha Pratim Talukdar, Joseph Reisinger, Marius Paşca, Deepak Ravichandran, Rahul Bhagat, and Fernando Pereira. Weakly-Supervised Acquisition of Labeled Class Instances using Graph Random Walks. In *Proc. of EMNLP-2008*, pp. 581–589, 2008.
9) Denyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schökopf. Learning with Local and Global Consistency. *NIPS*, Vol.16, pp. 321–328, 2004.
10) 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. 人工知能学会論文誌, Vol.23, No.3, pp. 217–225, 2008.
11) 萩原正人, 鈴木久美. 意味的類似度を利用した日本語クエリ書き換え. 言語処理学会第 15 回年次大会論文集, pp. 522–525, 2009.