原著論文 Original Paper

ラプラシアンラベル伝播による検索クリックスルーログからの意味カテゴリ獲得

Learning Semantic Categories from Search Clickthrough Logs Using Laplacian Label Propagation

小町 守 Mamoru Komachi
奈良先端科学技術大学院大学 Nara Institute of Science and Technology
mamoru-k@is.naist.jp, http://cl.naist.jp/~mamoru-k/

牧本 慎平 Shimpei Makimoto
ヤフー株式会社 Yahoo Japan Corporation
smakimot@yahoo-corp.jp

内海 慶 Kei Uchiumi
（同 上）
kuchiumi@yahoo-corp.jp

颯々野 学 Manabu Sassano
（同 上）
msassano@yahoo-corp.jp

keywords: search query logs, search clickthrough logs, semantic category, label propagation, semi-supervised learning

Summary

As the web grows larger, knowledge acquisition from the web has gained increasing attention. Web search logs are getting a lot more attention lately as a source of information for applications such as targeted advertisement and query suggestion. However, it may not be appropriate to use queries themselves because query strings are often too heterogeneous or inspecifiec to characterize the interests of the search user population. the web. Thus, we propose to use web clickthrough logs to learn semantic categories. We also explore a weakly-supervised label propagation method using graph Laplacian to alleviate the problem of semantic drift. Experimental results show that the proposed method greatly outperforms previous work using only web search query logs.

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．一般的なテキストと比較すると，**検索クエリ**は検索を行うユーザの関心を反映している [Silverstein 98] ため，ターゲット広告やクエリ展開のような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知識が重要だと考えられる．そして，自然言語処理においても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．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．

そこで，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する．本研究における**意味カテゴリ**とは，認知言語学のプロトタイプ理論に基づくカテゴリ論によるもので，古典的な必要十分条件によって定義されるカテゴリではなく，典型的なインスタンスとそれとの類似度によって特徴づけられる段階的な概念である [Rosch 75]．たとえば「鳥」という意味カテゴリの典型的なインスタンスはカラスやスズメであり，ダチョウやペンギンは周辺的なインスタンスである．このような段階性を持って他の意味カテゴリと区別されるような分類が「鳥」の意味カテゴリである．日本語語彙大系 [池原 97] における意味属性体系は，日本語の一般名詞・固有名詞・用言の意味的用法を約 3,000 の意味属性で体系化したものだが，本研究の対象とする意味カテゴリは検索ユーザや広告主の関心を反映したものとなるため，必ずしも体系化されていない．また，ウェブ検索では検索クエリとして入力される新しい固有名詞に対する意味カテゴリの情報が重要なのに対し，日本語語彙大系は出版以来更新されておらず（たとえば「携帯電話」という単語も登録されていない），本研究の目的には不適当である．

検索ログからの意味カテゴリ学習タスクとは，「シンガポール」のようないくつかの単語をシードとして検索クエリログ中に含まれる固有表現のリストから旅行カテゴ

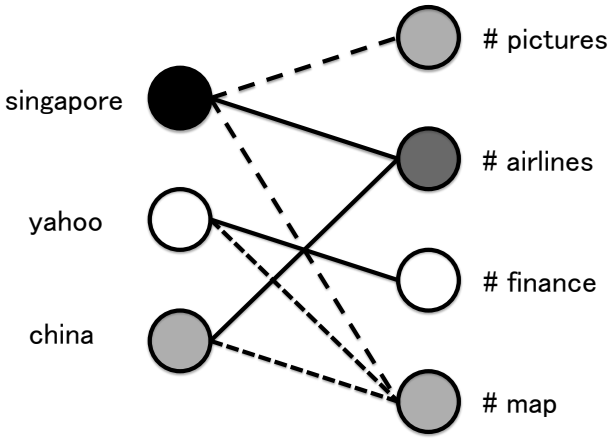

図 1 単語「singapore」から旅行カテゴリの単語を獲得する

リの他の単語を学習する，というタスクである．類似するタスクとしては，単語をシードとして意味カテゴリの名前を学習するタスク [Paşca 08] や意味カテゴリに属するパターンをシードとしてそのカテゴリに属する単語を学習するタスク [Hearst 92] がある。これらのタスクではコーパスを用いた手法が盛んに研究されており，その嚆矢となるブートストラップによるテキストからの情報抽出は [Hearst 92] によって提案された．ブートストラップとはシードインスタンス（単語）もしくはシードパターンからスタートし，反復的にパターン抽出・インスタンス獲得を繰り返すことで少数のデータから大規模なリソースを構築する手法である．

たとえば，検索クエリログをコーパスとして用いることを考える．シードインスタンス「シンガポール」を含む検索クエリとして,「シンガポール ビザ」や「シンガポール 写真」のようなクエリが取得できる．いま「シンガポール」がシードであることが分かっているので,「♯ビザ」や「♯写真」のようなクエリとの共起パターンを抽出できる．ここで，♯はクエリが元あった場所を示す．直感的に言うと，共起するクエリのパターンが似ている他の単語も「シンガポール」と似ている意味カテゴリであろう，という仮定に基づき，検索クエリログコーパスでこれらのパターンにマッチする単語を取得すれば，シードと同じカテゴリの単語を獲得できる．新しく獲得された単語をシードとして用いることでこのプロセスは反復的に繰り返すことができ，数個のシード単語を指定するだけでも大規模な単語辞書が構築できる，という利点がある．

その一例として，[小町 08] がある．彼らは大規模な日本語の検索クエリログを用い，*Espresso* [Pantel 06] に基づいて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．また，インスタンスとパターンのカテゴリへの所属度（スコア）を相互再帰的に計算するこれらのアルゴリズムは，2 部グラフ上の関連度の計算と見なすことができる [Komachi 08].

図 1 はシード単語として「Singapore」を用いて旅行カテゴリの単語を獲得するブートストラップの過程を示している．左側にはインスタンスのノード，右側にはパターンのノードがあるような 2 部グラフを考える．ノードの濃さが旅行カテゴリへの所属度を反映している．クエリ「singapore airlines」がコーパスに出現すれば，単語「singapore」からクエリ共起パターン「♯ airlines」へそれぞれの単語とクエリパターンの共起を重みとするエッジが張られる．リンク解析で用いられている手法を適用すれば，効率的にこのグラフ上でのシードインスタンスと任意のノード間の関連度が計算できる．

一方，ブートストラップには反復の際に多数のインスタンス集合と共起するパターン（**ジェネリックパターン**）を一度抽出してしまうと，それ以降シードと関連性の低いインスタンスを獲得する，という問題（**意味ドリフト**）が知られている．*Tchai* は検索クエリログに特化して意味ドリフトを抑えつつ意味カテゴリを学習することに成功し，ウェブ検索クエリログからの意味カテゴリ獲得において最も高い性能を示している．

しかし，この先行研究には以下の 2 点の問題があった．

リソースの問題 ウェブ検索クエリはバリエーションが広くかつ曖昧性も高く，必ずしも検索ユーザの意図を反映しているとは限らないため，パターンとしてウェブ検索クエリログを用いるのは適切ではない可能性がある．しかしながら，これまでの先行研究のほとんどは検索クエリログを用いている．

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* [Pantel 06] や，検索クエリログを対象とした *Tchai* [小町 08] では，使用するデータに合わせて，たとえばシードインスタンスの数，反復の停止条件，各反復で選択するインスタンスやパターンの数などの 8 個以上の変数をそれぞれ設定しなければならない．これらの変数の設定によってブートストラッピングは大きく性能が変化してしまう [Ng 03] 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．

そこで，本研究ではそれぞれの問題に対し，以下の 2 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

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．

[Joachims 02] は検索クリックスルーログを利用して検索エンジンのランキングを学習する手法を提案した。**検索クリックスルー**とは，検索の結果のランキングを見たユーザがクリックしたリンクのことである．同じクリックスルーを持つ検索クエリは同じ意図で検索された可能性が高い．従って，検索クリックスルーログは同義語や類義語の獲得タスクである意味カテゴリ学習において有

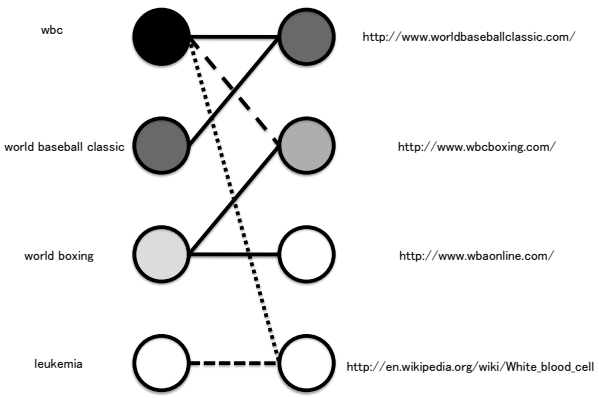

図 2 クリックスルーをバターンとして用いた意味カテゴリ学習

効であると考えられる．クリックスルーログは大量に入手可能であり，しかも低コストで保存できるという利点もあり，半教師あり学習の設定に適している．

② **ラプラシアンラベル伝播**による半教師あり学習

次に可搬性の問題について，提案手法はラプラシアンラベル伝播を用いて意味カテゴリ認識を行い，少ないパラメータ数（*Tchai* で 8 つあったパラメータを 1 つに減らす）で従来手法と同程度の性能を達成する．ラベル伝播手法などグラフに基づく手法は並列分散処理を用いることによる大規模化が容易であり，データスパースネスが大きな問題となるクリックスルーログの処理に適している．また，本研究はラプラシアンラベル伝播が正則化ラプラシアンカーネル [Smola 03] を用いたラベル伝播であることを示し，[Zhou 04] で用いられているヒューリスティクスを用いることなく，意味ドリフトの抑制に成功することを検証する．

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 [小町 08] があるが，意味カテゴリ獲得において検索クリックスルーログを活用し，グラフラプラシアンを用いたラベル伝播を適用した研究は，我々の知る限りこれが初めてのものである．

2. *Quetchup* アルゴリズム

本節では，グラフラプラシアンを用いたラベル伝播に基づく検索クリックスルーログから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*[*1] と名付けた．

2·1 クリックスルーパターンの使用

前述したように，検索クエリログはシードインスタンスと関連性の低いパターンを獲得する可能性がある．これは，検索クエリログにはたくさんのインスタンスと共起する**ジェネリックパターン** が含まれ，カテゴリを特徴付けるパターンやインスタンスの区別がつきにくい密なグラフを構成するためだと考えられる．

この問題を解消するために，我々は検索クエリログの代わりに検索クリックスルーログをパターンとして用いることを提案する．図 2 は検索クリックスルーログからどのようにパターンを構築するかを示している．図 1 と同様に，左側にインスタンスのノードを持ち，右側にパターンのノードを持つが，図 1 では検索クエリがパターンになっていたのに対し，図 2 では検索クリックスルーをパターンとして用いる．

クリックスルーをパターンとして用いる別の利点は，クエリをパターンとして用いるのと比較すると，カテゴリを特徴づけるパターンやインスタンスごとにまとまった，より疎なグラフを構成することである．クリックスルーログは検索ユーザがリンクをクリックするという行動により，暗黙的に曖昧性が解消されている．また，検索エンジンは重要度に従ったランキングを返すため，クリックスルーのバリエーションが少なく，それぞれの共起頻度は高くなる．これらの特徴により，クリックスルーログを用いたグラフのほうがより適切に単語とパターンの特徴を捉えていると考えられる．グラフの非連結成分が多いという問題は，データスパースネスの問題に起因するが，大規模なデータを用いることにより解消できる．並列分散計算によって大規模なデータを処理することで，この問題を解決できるというのは本研究の新たな知見である．実験結果は 3·2·4 節 で述べる．

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: 収束するまで $F(t+1) = \alpha AF(t) + (1-\alpha)F(0)$ を繰り返す

図 3 ラベル伝播アルゴリズム

2·2 ラプラシアンラベル伝播による半教師あり学習

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れ，また大規模化が容易であるという特徴がある．

[Zhou 04] によるラベル伝播アルゴリズムを図 3 に示した．まず，あらゆるインスタンスの集合を $\mathcal{X}$ と表すものとする．1 カテゴリの学習をする場合，$F(t)$ は $\mathcal{X}$ の要素数 $|\mathcal{X}|$ を次元数とするベクトルである．$F(t)$ の i 番目の次元の値は，$\mathcal{X}$ の i 番目のインスタンス x_i が対象のカテゴリに属する度合いを表す．すなわち，$F(t)$ は対象のカテゴリに対するスコアベクトルである．入力として与える $F(0)$ は，シードとして与えられるインスタン

*1 Query Term Chunk Processor

ス集合に x_i が含まれるとき，$F(0)$ の i 番目の次元の値を 1 とし，それ以外の次元の値は 0 とすることで作成する．アルゴリズムでは，このように作成された $F(0)$ を，行と列の次元がともに $|\mathcal{X}|$ のインスタンス類似度行列 A を用いて更新していくことで，最終的に収束した $F(t)$ が出力される．$F(t)$ は t ステップ終了時の $\mathcal{X}$ のスコアベクトルである．

また，2 カテゴリの学習をする場合，$F(t)$ は $\mathcal{X}$ の要素数 $|\mathcal{X}|$ を次元数とするベクトルと扱うことができる．$F(0)$ はシードとして与えられるインスタンスに対応する次元の値を，2 つのクラスに応じて 1 または -1 とする．

3 つ以上の n 個のカテゴリを学習する場合は $F(t)$ を $|\mathcal{X}|$ 行 n 列の行列とする．j 番目の列ベクトルが各カテゴリに対応したスコアベクトルとなる．最終的に得られた $F(t)$ において，あるインスタンスがどのクラスに属するかは，どの列のスコアベクトルの値が最も大きな値を持っているかで判定する．2 カテゴリの場合にもこの手法は適用可能であり，また，1 カテゴリの場合はこの手法の $n=1$ とした場合である．

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．

しかしながら，このアルゴリズムは類似度行列の作成方法によっては意味ドリフトが起きるという問題点がある．特に類似度行列 A をインスタンス・パターンの共起行列 W を用いて計算し，$A=W^TW$ とした場合，図 3 で示されるアルゴリズムの出力は Kleinberg の HITS [Kleinberg 99] における権威度ベクトルと一致するため，シードによらないスコアベクトルを返すことが示されている [伊藤 04]．この現象は HITS においてはトピックドリフトとして知られており，ブートストラップでは意味ドリフトと呼ばれていた問題であることが [Komachi 08] によって示された．[Zhou 04] では類似度行列のスコア $A_{ii}=0$ にすることでこの問題に対処しているが，自己類似度を 0 にするのはヒューリスティクスであり，理論的な裏付けがあるものではなかった．

そこで，我々は類似度行列として正規化ラプラシアンを用いた手法を提案する．[Zhou 04] との違いは類似度行列の作成方法とグラフラプラシアンの適用である．グラフラプラシアンはグラフ中の自己類似度の重みを減じる効果があるため，ジェネリックパターンに高い重みを付与することがなく，意味ドリフトが起こりにくいという利点がある．

正規化ラプラシアンを用いたラベル伝播手法を図 4 に示す．D は $D_{ii}=\sum_j N_{ij}$ で定まる行列 N の次数対角行列であり，インスタンス・パターン共起行列 W の W_{ij} 要素は行で正規化する．図 3 との違いは類似度行列のスムージングとグラフラプラシアンの適用によって意味ドリフトに対処することである．

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: 正規化ラプラシアン行列 $L=I-D^{-1/2}AD^{-1/2}$ を作成する
2: 収束するまで $F(t+1)=\alpha(-L)F(t)+(1-\alpha)F(0)$ を繰り返す

図 4 ラプラシアンラベル伝播アルゴリズム

反復的アルゴリズムにより，

$$F(t)=(\alpha(-L))^{t-1}F(0)+(1-\alpha)\sum_{i=0}^{t-1}(\alpha(-L))^iF(0)$$

$0<\alpha<1$ なので，$(-L)$ の固有値は $[-1,1]$ であり，$\lim_{t\to\infty}(\alpha(-L))^{t-1}=0$ となるため，

$$\lim_{t\to\infty}\sum_{i=0}^{t-1}(\alpha(-L))^i=(I-\alpha(-L))^{-1}=(I+\alpha L)^{-1}$$

従って

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

図 5 $F(t)$ が $F^*=(1-\alpha)(I+\alpha L)^{-1}F(0)$ に収束することの証明

2·3 ラプラシアンラベル伝播と正則化ラプラシアンカーネルの関係

ラベル伝播などのグラフに基づく半教師あり手法は，各ステップでラベルの重み付き投票をすることによってノードのスコアを求め，最終的な分類を行なうことに相当するが，ラプラシアンラベル伝播の重み付けは，リンク解析分野でトピックドリフトが起きにくいことが知られている正則化ラプラシアンカーネル [Smola 03] によるものであることを示す．

まず，列 $F(t)$ は $F^*=(1-\alpha)(I+\alpha L)^{-1}F(0)$ に収束することを図 5 で証明する．

ここで，分類においては $(1-\alpha)$ は無視できるので，

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

となるが，これは [Smola 03] の正則化ラプラシアンカーネルとまさに同一の式であり，ラプラシアンラベル伝播は実質的には正則化ラプラシアンカーネルによって重み付けしたラベルの投票によって最終的な分類を行なう．

正則化ラプラシアンカーネルはリンク解析において関連度を測る尺度として広く使われ，その有効性が確認されている [伊藤 04]．また，自然言語処理においても語義曖昧性解消タスクで有効であることが示されている [Komachi

08]．よって，意味カテゴリ学習タスクでも正則化ラプラシアンカーネルが有効であることが期待され，実際有効であることを後に実験で示す．

2·4 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．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．そのため，大規模データに対してはいくつか計算上の工夫が必要である．

§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，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 万程度なので，非現実的である．そこで，類似度行列 A を $A = W^T W$ 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，現実的なサイズに落とし込むことができる．また，W は 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できる．

§2 近似による計算量の削減

次に計算量について述べる．(1) によって計算されるラプラシアンラベル伝播計算は固有値計算を伴うため，ナイーブな計算量は $O(n^3)$ 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．特に $t = 0$ のときはシード単語のラベルのみを用いることに対応し，$t = 1$ のときはシード単語と共起するパターンと共起する単語にラベルを伝播する．

このようにラベル伝播を近似すると，2 つの利点がある．1 つはすでに述べたように計算量が $O(n^2 t)$ ($A = W^T W$ によって計算した場合 $O(npt)$) で抑えられるということと，もう 1 つは並列分散計算が容易なことである．具体的には，$F(t+1)$ を求める際 $\alpha(-L)F(t)$ の計算量が $O(np)$ であるが，MapReduce を適用することでこの計算を並列化できる．ディスクアクセスの局所性を利用し L はローカルディスクに保存することで効率的に反復計算を行える．また，Lanczos 法など反復計算によって固有値計算を行う場合，解が収束するまでにかかる時間の見積もりがしにくいというのは実用上大きな問題であるが，近似することによって見積もり可能な時間内で一定の精度の解を求められる．

3. 実　　験

本節では先行研究 *Tchai* [小町 08] と提案手法 *Quetchup* の比較実験について述べる．

3·1 実 験 設 定

§1 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．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．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．

§2 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．計算時間の短縮と行列のサイズの圧縮のため，クリックされた回数が 200 回以下のアドレスは削除し，共起するクエリが 1 つだけのアドレスも削除した．

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」（♯ は単語の入る位置を示す）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・パターン共起行列 W の (i, j) 要素 W_{ij} は行によって正規化する．

$$W_{ij} = \frac{|x_i, p_j|}{\sum_k |x_i, p_k|}$$

ここで x_i はインスタンス，p_j はパターン，$|x_i, p_j|$ はインスタンス x_i とパターン p_j の共起回数である．

§3 対象カテゴリ

今回のタスクは [小町 08] に倣い，「旅行」カテゴリと「金融サービス」カテゴリの 2 カテゴリを対象とした．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．

ある単語があるカテゴリに含まれるか否かはその単語がそのカテゴリを特徴づける語と共起しているかによって判定した．検索クエリログ中で 2 単語目に出現する頻出パターンを見るほか，その単語でウェブ検索を行った結果も参考にした．本来意味カテゴリは段階性を持つものだが，本研究ではタグ付けの簡易化のため，そのカテゴリであるとき 1，そうでないとき 0 を付与した．

また，クエリに表記揺れや綴り誤りがある場合，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，現実的な状況を想定し，代表表記に復元したうえで意味カテゴリを付与した．1 クエリに複数単語が含ま

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
金融サービス	みずほ銀行，三井住友銀行，jcb，新生銀行，野村證券

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．

§4 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．

t ステップ近似を用いたラベル伝播は，インスタンスと直接共起しないパターンを用いてスムージングを行うことに相当する．行列が比較的密であれば，1-2 回のステップで停止することでも，意味ドリフトを防ぎつつラベル伝播を行うことができる．しかし，予備実験の結果，最低でも 5 回程度のステップを繰り返さないと，再現率が非常に低くなることが判明したため，反復回数を 10 回とした．これは，クリックスルーログが非常に疎であるためだと考えられる．*Tchai* は [小町 08] と同じパラメータで実行し，1 回当たり 10 インスタンスずつ 10 回反復し，100 インスタンスを取得した．同様に *Quetchup* の反復回数は 10 回，パラメータ α は 0.0001 とした．MapReduce の実装はオープンソースの Hadoop [*2] を用いた．

§5 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 [Pantel 04] を用いた．精度は（システムが出力した正解の個数）/（システムが出力したインスタンスの個数）である．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．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．

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

$R_{A|B}$ 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 を分子と分母で相殺することで，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 $|A|, |B|$ から相対再現率を求めることができる．

3·2 実 験 結 果

§1 検索クリックスルーログの有効性

図 6 から 図 9 は検索クリックスルーログの効果を示すために 3 つのシステムにおける精度と相対再現率をプロットしたものである．

図 6 と図 7 からは，旅行カテゴリ・金融カテゴリ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ことが分かった．*Quetchup*（**クエリ**）を見ると，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ていた．実際に獲得されるクエリを分析したところ，アダルト情報に関する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，図 8 と図 9 によると，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で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既存手法より優れていることが分かる．

図 6–図 8 と図 7–図 9 を比較すると，旅行カテゴリと金融カテゴリではクエリを用いたシステムの振る舞いに違いが見られるが，これはそれぞれのカテゴリに含まれる正解の数の違いによるものだと考えられる．金融カテゴリは閉じた集合であり，効果的なクエリパターンは少数しかないが，旅行カテゴリに含まれる単語は膨大であり，これらの単語にマッチするパターンが多数ある．従って，意味ドリフトが起きた場合，パターンが少数の金融カテゴリでは大きな影響が現れるが，旅行カテゴリでは影響は比較的軽微である．

§2 クリックスルーとクエリの混合

3·2·1 節においてクリックスルーログとクエリログ両方を用いた行列を作成する際，まずそれぞれについて正規化し，パラメータ $\beta \in [0,1]$ によって混合した．

$$W_{mix} = \beta W_{click} + (1-\beta) W_{query}$$

ここで W_{mix} は行の合計が 1 になるように正規化されたインスタンス・（クリックスルーとクエリからなる）パターン行列である．クリックスルーとクエリパターン両方を用いた類似度行列 A_{mix} は

$$A_{mix} = W_{mix}^T W_{mix}$$

により与えられる．

クリックスルーとクエリログの割合は，β を 0 から 1 まで 0.2 刻みで変化させた．$\beta = 0.2$ がクリックスルー対クエリの比率 2:8 に相当する．また，*Quetchup*（**クリック**）と *Tchai* の相対再現率は，*Quetchup*（**クエリ**）に対して計算した．

図 10 と図 11 から，検索クリックスルーの割合を増やすと精度と相対再現率が上昇することが確認され，検索

*2 http://hadoop.apache.org/

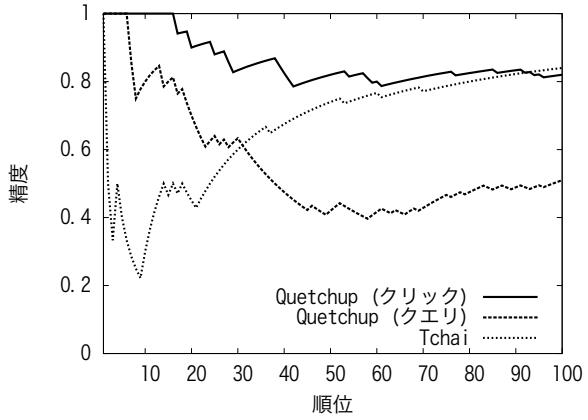

図 6 旅行カテゴリにおける精度

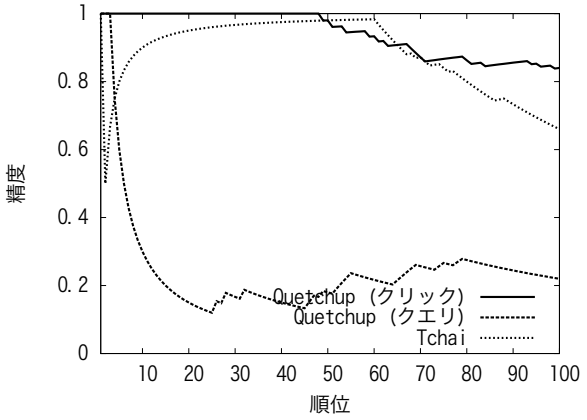

図 7 金融カテゴリにおける精度

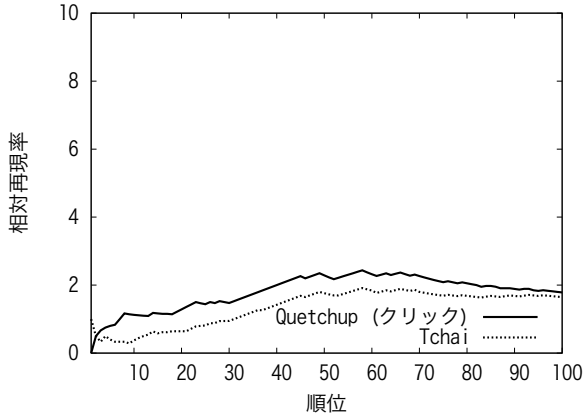

図 8 旅行カテゴリにおける相対再現率

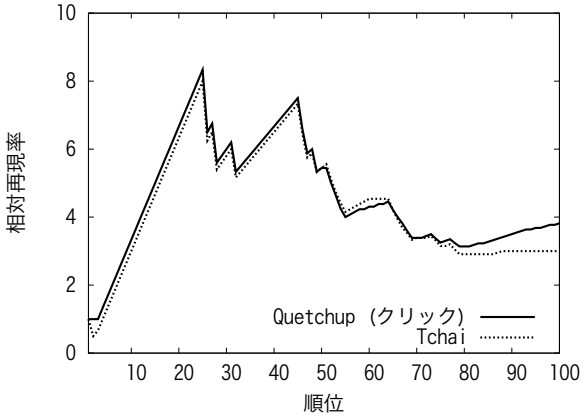

図 9 金融カテゴリにおける相対再現率

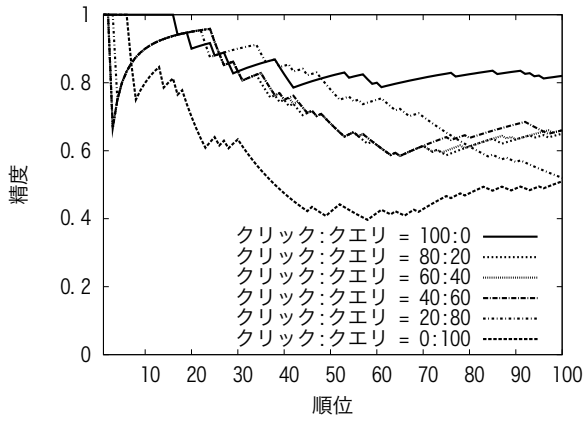

図 10 クリック対クエリの割合を変化させたときの *Quetchup* の精度

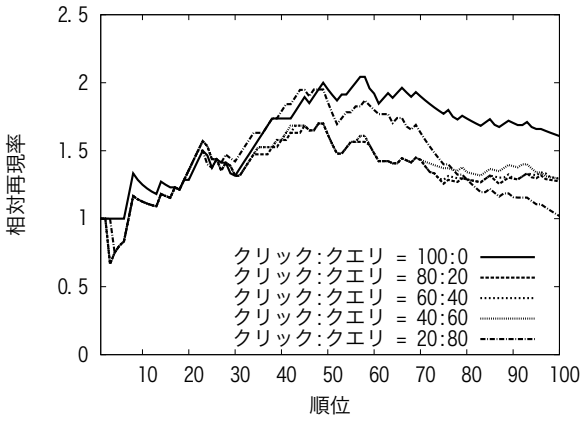

図 11 クリック対クエリの割合を変化させたときの *Quetchup* の相対再現率

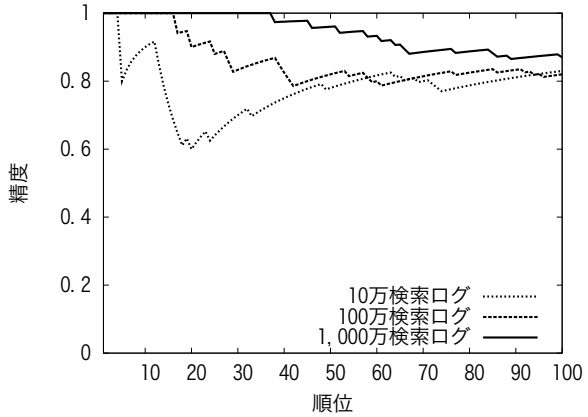

図 12 検索クリックスルーログの量を変化させたときの精度

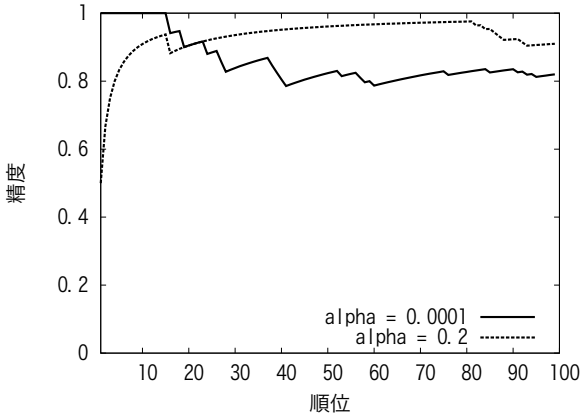

図 13 パラメータ α を変化させたときの精度

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．図 11 の実線「クリック:クエリ=100:0」を見ると，相対再現率はクエリのみを用いたシステムに比べ 1.5-2 倍であり，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．

§3 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，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．とはいえ，これは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．

一方，クエリを用いると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，意味ドリフトが起きてしまう．

表 3 は獲得されたインスタンスの分布を調べるため，異なり頻度上位 1,000 万件のクリックスルーログを用いて上位 10,000 件のクエリを獲得し，そのうち 100 件をランダムサンプリングして人手で分類したものである．

もっとも数が多いクエリは交通機関であり，全体の半分以上を占めている．これは，シードとして用いたクエリ 5 つ中 3 つまでが jal, ana, jr と交通機関に関するシードであったため，鉄道や交通会社に関するクエリを抽出する傾向が強いことを示している．

次に多いのは宿泊施設および旅行情報のクエリであり，これらのサブカテゴリに対するシードクエリは与えていないものの，旅行に関する固有表現の抽出に成功している．

また，全体のうち 20%のクエリは固有表現が含まれないクエリであったが，表 3 の「それ以外」の例に示したように，そのうち 1/4 は旅行に関するナビゲーショナルクエリ [*3]であった．さらに 10,000 件のクエリを獲得しても旅行に関係のないクエリは全体の 2 割程度であり，同義語や表記揺れ以外に，高い精度で意味カテゴリ学習が可能であることが分かる．

§4 データサイズを増減したときの性能比較

図 12 は *Quetchup* アルゴリズムに対し，データ量によるふるまいの違いを見るため，データサイズを検索ログ頻度上位 10 万件，100 万件，1,000 万件と変えて精度を比較した結果である．

用いるデータ量が少ない場合に比べ，データ量を増やすと精度が向上することが示された．これは，クリックスルーログはクエリログに比べてスパースで，データ量が増えれば増えるほどグラフが密になり，シード単語から到達できるノードが増えるためであると考えられる．

§5 パラメータ α を変化させたときの性能比較

図 13 は *Quetchup* アルゴリズムに対し，ラベルありデータとラベルなしデータのどちらをより重視するかをコントロールするパラメータ α を変化させたときの精度の変化を示す．α は 0 から 1 まで 0.2 刻みで変化させて実験したが，ほぼ 0.2 の結果と一致するので，簡単のために 0.2 の場合のみを図示した．α が大きい場合はグラフ構造を重視するため意味ドリフトが起き，α が小さい場合はシードを重視するために意味ドリフトは起きないことが予想される．

しかしながら，実験の結果，α が大きい場合のほうが α が小さい場合よりむしろ精度が高いことが確認された．これは，クリックスルーログから作成されるグラフの場合，クエリログから作成されるグラフと比べ，非常に疎なグラフとなっており，またユーザが実際にクリックしたという情報を用いてグラフを作成しているため，グラフ中での重要度が旅行カテゴリでの重要度を反映するものになるためだと考えられる．

実際，シードとして用いたクエリの $\alpha = 0.8$ における順位はそれぞれ jr (43), じゃらん（378）, ana (755), jal (904), his (1362) となっており，シードが必ずしも上位にランクされていないことから，意味ドリフトは起きているものの，旅行カテゴリの単語が獲得されていることが分かる．これは，必ずしも今回用いたシード単語が今回のデータセットに対して最適のシードではなかったということを示唆している．

4. 関 連 研 究

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 [Paşca 07a, Paşca 07b]．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．

[Xu 09] のタスクは，検索クリックスルーログを用いて固有表現の意味カテゴリ学習を行う点で，我々のタスクと共通している．しかしながら，彼らはクリックスルーデータを LDA [Blei 03] によってモデル化し，シードインスタンスによって定まる意味カテゴリのラベルを教師データとして，確率モデルによって意味カテゴリを推定するため，我々と手法が異なる．

同様に，[Li 08] は検索クリックスルーログを用いて検索意図を学習するタスクである．検索クリックスルーログから作成したグラフを正規化してラベル伝播を行うという手法は共通しているが，彼女らはグラフラプラシア

*3 特定のリンク先に決まっているクエリ [Broder 02]．たとえば http://www.youtube.com/に対する「Youtube」．

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）**	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）**	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（1 ステップ目）	jtb, 新幹線, 航空券, 全日空, 飛行機, 格安, 国内線, 旅行, 高速バス, jr 東日本	♯ キャンセル料, ♯ キャンセル, ♯ 早割り, ♯ 北海道, ♯ キャビンアテンダント, ♯ グランドスタッフ, ♯ 陸マイラー, ♯ スカイメート, ♯ 機内販売, ♯ 介護割引
Tchai（10 ステップ目）	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

表 3 獲得されたインスタンスのランダムサンプリング

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	ホテルビーナス, リーガロイヤルホテル大阪, www.route-inn.co.jp, ホテル京阪ユニバーサル・シティ, 札幌全日空ホテル
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, youtubu 海外ドラマ, 法務部採用, おくりびと, 社会人野球

ンを使っておらず，タスクも異なっている．

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 [Paşca 08, Talukdar 08]．[Talukdar 08] 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リックスルーログを使う点と，グラフの作成方法が異なる．

Tchai [小町 08] はブートストラップ手法の一種である．彼らの手法は高い精度で意味カテゴリ学習を行うことができるが，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．また，検索クリックスルーを使用せず，検索クエリログのみを用いている．

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る．一方，実用的にはシード単語の選択も重要である．[Vyas 09] で提案されているような尺度によって，作業者のシード選択コストの削減を考える予定である．

クリックスルーログをはじめ，ユーザの入力データには様々な利用方法が考えられる．たとえば，ユーザが同一セッション内でどのようなクエリを入力したかのログであるセッションログから意味クラスの属性を抽出することも可能であろう．自然言語処理ではタグつきデータの作成コストがしばしば問題になるが，大規模なユーザの入力データを用いた手法は，ノイズを含むがデータ作成コストを大幅に削減できる．今後もユーザの入力データの効果的な活用方法を研究していきたい．

謝　　辞

有益な示唆をくださった匿名の査読者の方々と WS-LDA に関する引用の間違いを指摘してくださった坪坂正志氏にお礼申し上げる．ヤフー株式会社北岸郁雄氏には共同研究を進めるに当たってお世話になった．また，小間基裕氏からは検索ログデータの利用，町永圭吾氏からはクリックスルーログの前処理プログラムでご協力いただいた．合わせて感謝する．

◇ 参 考 文 献 ◇

[Blei 03] Blei, D., Ng, A., and Jordan, M.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)

[Broder 02] Broder, A.: A Taxonomy of Web Search, *ACM SIGIR Forum*, Vol. 36, No. 2, pp. 3–10 (2002)

[Hearst 92] Hearst, M.: Automatic Acquisition of Hyponyms from Large Text Corpora, in *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pp. 539–545 (1992)

[Joachims 02] Joachims, T.: Optimizing Search Engines Using Clickthrough Data, in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 133–142 (2002)

[Kleinberg 99] Kleinberg, J.: Authoritative Sources in a Hyperlinked Environment, *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632 (1999)

[Komachi 08] Komachi, M., Kudo, T., Shimbo, M., and Matsumoto, Y.: Graph-based Analysis of Semantic Drift in Espresso-like Bootstrapping Algorithms, in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 1010–1019 (2008)

[Li 08] Li, X., Wang, Y.-Y., and Acero, A.: Learning Query Intent from Regularized Click Graphs, in *Proceedings of SIGIR'08: the 31st Annual ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 339–346 (2008)

[Ng 03] Ng, V. and Cardie, C.: Weakly Supervised Natural Language Learning Without Redundant Views, in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 94–101 (2003)

[Paşca 07a] Paşca, M.: Organizing and Searching the World Wide Web of Fact — Step Two: Harnessing the Wisdom of the Crowds, in *Proceedings of the 16th International World Wide Web Conference (WWW-07)*, pp. 101–110 (2007)

[Paşca 07b] Paşca, M. and Durme, B. V.: What You Seek is What You Get: Extraction of Class Attributes from Query Logs, in *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, pp. 2832–2837 (2007)

[Paşca 08] Paşca, M. and Durme, B. V.: Weakly-Supervised Acquisition of Open-Domain Classes and Class Attributes from Web Documents and Query Logs, in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-2008)*, pp. 19–27 (2008)

[Pantel 04] Pantel, P. and Ravichandran, D.: Automatically Labeling Semantic Classes, in *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-04)*, pp. 321–328 (2004)

[Pantel 06] Pantel, P. and Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pp. 113–120 (2006)

[Rosch 75] Rosch, E.: Cognitive Representation of Semantic Categories, *Journal of Experimental Psychology*, Vol. 104, pp. 192–233 (1975)

[Silverstein 98] Silverstein, C., Henzinger, M., Marais, H., and Moricz, M.: *Analysis of a Very Large AltaVista Query Log*, Digital SRC Technical Note 1998-014 (1998)

[Smola 03] Smola, A. J. and Kondor, R. I.: Kernels and Regularization of Graphs, in *Proceedings of the 16th Annual Conference on Learning Theory*, pp. 144–158 (2003)

[Talukdar 08] Talukdar, P. P., Reisinger, J., Paşca, M., Ravichandran, D., Bhagat, R., and Pereira, F.: Weakly-Supervised Acquisition of Labeled Class Instances using Graph Random Walks, in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 581–589 (2008)

[Vyas 09] Vyas, V., Pantel, P., and Crestan, E.: Helping Editors Choose Better Seed Sets for Entity Set Expansion, in *Proceedings of ACM Conference on Information and Knowledge Management (CIKM-2009)*, pp. 225–234 (2009)

[Xu 09] Xu, G., Yang, S., and Li, H.: Named Entity Mining from Click-Through Log Using Weakly Supervised Latent Dirichlet Allocation, in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1365–1373 (2009)

[Zhou 04] Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schökopf, B.: Learning with Local and Global Consistency, *Advances in Neural Information Processing Systems*, Vol. 16, pp. 321–328 (2004)

[伊藤 04] 伊藤 敬彦, 新保 仁, 工藤 拓, 松本 裕治：カーネル法による計量書誌尺度の統一的解釈, 人工知能学会論文誌, Vol. 19, No. 6 SP-C, pp. 530–539 (2004)

[小町 08] 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)

[池原 97] 池原 悟, 宮崎 正弘, 白井 諭, 横尾 昭男, 中岩 浩巳, 小倉 健太郎, 大山 芳文, 林 良彦（編）：日本語語彙大系, 岩波書店 (1997)

〔担当委員：松 尾 豊〕

2009 年 6 月 10 日　受理

著 者 紹 介

小町　守（学生会員）

2005 年東京大学教養学部基礎科学科科学史・科学哲学分科卒．2007 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．同年後期課程に進学．修士（工学）．日本学術振興会特別研究員（DC2）．大規模コーパスを用いた意味解析に関心がある．言語処理学会同第 14 回年次大会最優秀発表賞受賞．言語処理学会，情報処理学会，ACL 各会員．

牧本　慎平

2006 年広島大学総合科学部総合科学科卒業．2008 年奈良先端科学技術大学院大学情報科学研究科情報処理学専攻博士前期課程修了．同年ヤフー株式会社入社．現在同社 R&D 統括本部プラットフォーム開発本部勤務．ウェブ検索に関する自然言語処理，データマイニングの研究開発に従事．言語処理学会，情報処理学会各会員．修士（工学）．

内海　慶

2004 年図書館情報大学図書館情報学科卒．2006 年筑波大学大学院図書館情報メディア研究科博士前期課程修了．同年 4 月ヤフー株式会社入社．2009 年現在同社在職中．自然言語処理の研究開発に従事．修士（情報学）．言語処理学会会員．

颯々野　学

1991 年京都大学工学部電気工学第二学科卒業．同年より富士通研究所研究員．1999 年より 1 年間，米国ジョンズ・ホプキンス大学客員研究員．2006 年よりヤフー株式会社勤務．自然言語処理の研究に従事．2008 年京都大学大学院情報学研究科知能情報学専攻博士後期課程修了．博士（情報学）．言語処理学会，情報処理学会，ACL 各会員．