DBSJ Journal, Vol.12, No.1
June 2013

語の出現の偏りに基づく新たな隠語の発見

Discovering New Jargons Based on Skew of Word Appearance Distribution

大西 洋♥　田島 敬史◆

Hiroshi OHNISHI　Keishi TAJIMA

犯罪行為に関する情報や有害情報を，一般のユーザや監視機関に対しては隠蔽しつつ，一部のユーザ間で伝達するために，**Web** 上では日々新たな隠語が生まれており，こうした隠語の早期発見は犯罪防止や有害情報のフィルタリングに有用である．本研究では，隠語がアンダーグラウンド系掲示板に偏って現れることを用いて，新たな隠語を発見する手法を提案する．提案手法では，アンダーグラウンド系掲示板の新着記事から隠語候補を取得し，その語による **Web** 検索結果とその記事をクラスタリングして，その記事が隠語が偏在するクラスタに入るかにより，語が隠語か判定する．また，隠語候補を探すべきアンダーグラウンド系掲示板も日々変わっていくため，新たな隠語の発見と新たなアンダーグラウンド系掲示板の発見を並行して行う．

New jargons are born day by day for exchange of illegal or harmful information on the Web without the exposure to ordinary people or the authorities. Discovery of them is useful for crime prevention and filtering. In this paper, we propose a method of discovering new jargons based on the skew of their appearance distribution to underground BBSs. Our method retrieves new postings from underground BBSs, also retrieves Web pages by using each word in each posting as a query, and then classifies the retrieved pages and the posting into clusters corresponding to the usages of the word. We judge the word to be a jargon if the posting is classified into a cluster containing many jargons. Because underground BBSs also appear and disappear day by day, we find new underground BBSs in parallel to the discovery of new jargons.

1. はじめに

隠語とは,「社会的集団内部の秘密保持・隠蔽のために内部の人間だけが分かるように造られ使用されることば」[11] であり，例えば薬物取引の場で「アイス」は覚醒剤を，出会い系の掲示板 (BBS) で「サポ」は援助交際を指す．米川 [11] が挙げた以下の隠語の社会的機能は，隠語の本質を捉える上で重要である (強調筆者).

- 所属集団の**秘密を保持**する機能
- 秘密保持により，**仲間意識**や**連帯意識**を強化する機能
- 集団のアイデンティティを確認し，**他の集団と区別**する機能
- 他の集団に対して自己を**誇示・自慢**する機能

隠語の機能のうち所属集団の秘密を保持する機能は，集団外の他者がその語を見ても内容が分からないようにし，情報の隠蔽を図るものである．特に，隠語を用いる社会集団が反社会的集団の場合，この「他者」は警察その他の治安組織である．薬物の販売や援助交際の勧誘などのアンダーグラウンドでの取引では，治安組織の捜査の目を逃れるために，露骨に違法行為を表す語ではなく隠語が用いられる．また，隠語は公知になるとその意義を失ってしまうため，絶えず新たな隠語が生み出されてきた．

近年，犯罪行為に関する情報や有害情報の伝達において，アンダーグラウンドな Web ページ，特にアンダーグラウンド系 BBS を介在する事例が増加している．BBS は誰でも手軽に見られることもあり，こうした書き込みを安全に行うため，以前にもまして多様な隠語が生み出されている．

このような現状を放置すれば，Web 上に有害情報が蔓延し犯罪の増加に繋がりうる．そこで本研究では，アンダーグラウンド系 BBS で新たに生まれた隠語の早期発見を目的とする．これにより，治安組織やフィルタリング事業者が手作業で隠語を探す負担が軽減され，犯罪抑止や効率的なフィルタリングが可能になる．

上述のように，本研究では，新たな隠語の発見場所として BBS に着目する．これは，一方的な情報発信しかなされない静的な Web ページより，利用者間でコミュニティが形成される BBS の方が新語が生まれやすいと考えられるからである．また，多くの閲覧者がいる BBS は，薬物の密売人などにとって恰好の宣伝の場だが，同時に人目につきやすいため危険性も大きい．そのため，秘密保持のため隠語を用いた投稿がなされる確率が高いといえる．

逆に言えば，閲覧者が少ないページや閲覧者が固定されたページではわざわざ隠語を使う必要があまりないため，隠語が用いられる Web ページは，一定数以上の閲覧者がいるページだともいえる．従って，隠語は Web 全体のうちでもアンダーグラウンド系 BBS に偏在していると考えられる．

また，例えば薬物に関する話題の場合，秘密保持のためには薬物の名前だけでなく，薬物摂取に用いる道具や取引用語など，周辺に現れる語も隠語化する必要がある．薬物の名前を隠語化しても，注射器や販売といった語をそのまま用いると，こうした語から薬物の話題であることが容易に推測される．そこで注射器を「ジェクター」，販売を「手押し」と隠語化し，他者による文意の推測を困難にする．このような面からも，隠語は他の隠語と共起して偏在して現れると考えられる．

提案手法では，BBS から隠語を抽出する際に，この隠語の偏在性を用いる．まず，アンダーグラウンド系 BBS への投稿中の各語とその周辺にある語をクエリとして Web を検索し，検索結果のページ集合に元の投稿を加えたものを語の用法別にクラスタリングする．語を隠語として用いているクラスタがあれば，そのクラスタにはアンダーグラウンドなページが偏って含まれるはずである．従って，元の投稿を含むクラスタがアンダーグラウンドなページに偏ったクラスタならば，その語を隠語と判定する．

秘密保持のため絶えず新たな隠語が作られるように，アンダーグラウンド系 BBS も新たに作られては，治安組織の監視などにより短期間で使われなくなる．従って，新たな隠語の早期発見には，最新のアンダーグラウンド系 BBS の発見も必要である．そこで本研究では，最新の隠語を含むクエリで Web 検索を行い，その検索結果で得られたページがアンダーグラウンドなページであるか，また，BBS であるか判定することで，最新のアンダーグラウンド系 BBS を発見する手法も提案する．

上述の通り本研究では，最新の隠語の発見に最新のアンダーグラウンド系 BBS を用い，最新のアンダーグラウンド系 BBS の発見に最新の隠語を用いる．そこで，既知の隠語の辞書と既知のアンダーグラウンド系 BBS の辞書から始め，両者を並行して更新することで，興廃の激しい隠語やアンダーグラウンド系 BBS の恒常的な発見を可能にする．

♥ 京都大学大学院情報学研究科修士課程　ohnishi@dl.kuis.kyoto-u.ac.jp
◆ 正会員　京都大学大学院情報学研究科　tajima@i.kyoto-u.ac.jp

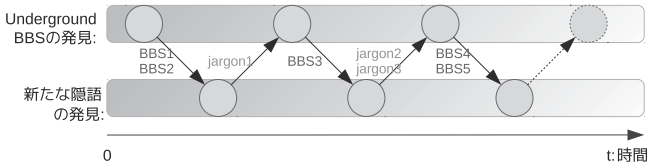

図 1: 提案手法の概念図

Fig 1. Concept of our method

2. 関連研究

多くの隠語は，既知の有害語と同義語の関係にある．山本ら [6] は，未知語が与えられた時に，その語と前後に現れる語が類似する語を発見することで，その語の同義語，関連語を発見する手法を提案した．そこで，BBS 中の未知語に対して同義語を求め，既知の有害語と同義語である未知語を隠語と判定する方法が考えられる．しかし，隠語には「アイス」「葉っぱ」等の一般語が用いられることも多いため，未知語とはならない．また，隠語が使用される際には，その隠語の前後にも明示的に違法行為を表す語は現れないことが多いため，これらの「アイス」などの語に対して前後の語を用いて同義語を発見しても,「覚醒剤」を抽出するのは難しい．同様の理由から，反対向きに「覚醒剤」などの既知の有害語の同義語を探しても，隠語の発見は困難である．既知の隠語が存在する語に対しては，その既知の隠語の同義語を発見することも考えられるが，あらゆる既知の隠語に対して手法を適用しない限り，指定した特定の隠語に対する新語しか発見できない．さらに，発見された同義語が隠語であるかの判定も別途必要となる．

那 [8] は，既知の新語と収集した Web ページに潜在意味解析 (LSA) を適用した結果の類似度を用いて，新語と類似する既存の語を発見する手法を提案した．この手法も，既知の隠語の同義語の発見にも応用可能と考えられ，また，前後の語のみを用いる手法よりも，われわれの手法に近いが，やはり，指定した語の新語しか発見できず，また，隠語であるかの判定が別途必要である．

橋本ら [7] は，与えられた語の周辺語を用いて語の隠語判定を行い，周辺語を用いない隠語判定より精度が上がることを示した．しかし，この手法では，例えば隠語「草」の検出には,「指定駅で草を手渡します」のように，その語を隠語として用いている文例を予め人手で集める必要がある．そのため，そうした文例が無い状態では隠語判定ができない．

また，これらの研究では一般の Web ページを探索の対象としている．一方，本研究では，隠語が偏在するアンダーグラウンド系 BBS を発見する手法を提案することにより，隠語かどうかの判定を可能とするとともに，全ての Web ページを対象とするよりも効率的に隠語を発見することを可能としている．

情報フィルタリングの面からも多くの先行研究がある．松葉ら [9] は学校非公式サイトでの有害情報の伝達を規制するため，有害語を含む BBS の投稿をサポートベクタマシン (SVM) に学習させ，投稿をフィルタリングする手法を提案した．しかし，有害語は人手で集めるため，未知の有害語への対応は研究課題としている．

3. 提案手法

本研究では，事前に人手で隠語の用例を収集せずとも，自動で Web 上から新たな隠語を発見できる手法を提案する．提案手法は，既知の隠語を用いて新たな他の意味の隠語も発見するスノーボールサンプリング型の手法で，図 1 のように動作する．図中の丸は動作中の部分を，矢印は各部間のデータのやりとりを表す．

1 章で述べた通り，本研究はアンダーグラウンド系 BBS への投稿に着目している．そのため提案手法では，アンダーグラウンド系 BBS の発見部と新たな隠語の発見部が存在し，これらが並列に

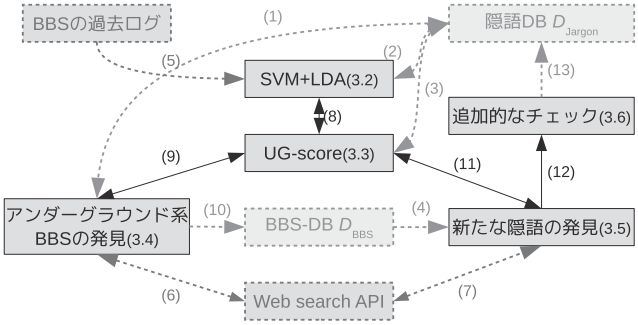

図 2: 提案手法の全体図

Fig 2. Overview of our method

動く．図 1 ではまずアンダーグラウンド系 BBS として「BBS1」「BBS2」を発見し，アンダーグラウンド系 BBS の辞書 (BBS-DB) に加える．次に BBS-DB 中の BBS への新着投稿から，隠語「jargon1」を発見し，隠語データベース (隠語 DB) に加える．続いて，隠語 DB 中の隠語を用いてアンダーグラウンド系 BBS を発見し，以降も同様に動作する．この手法により，隠語やアンダーグラウンド系 BBS の興廃に柔軟に対応できる．

より詳細な提案手法の全体図を図 2 に示す．図中の四角は手法の構成要素を表し，構成要素に括弧書きで付加した数字は，その構成要素の詳細を述べた節の番号を表す．実線の四角は本研究での実装部を，点線の四角は提案手法で利用または作成した，辞書もしくは Web 上の資源を表す．また実線の矢印は提案手法での処理の流れを，点線の矢印はデータの入力または修正を表す．

3.1 提案手法の概要

本節では，図 2 を参照しつつ，提案手法の概要を述べる．

1 章で述べた通り，本研究では隠語発見のため，まずアンダーグラウンド系 BBS を発見することが必要となる．アンダーグラウンド系 BBS を自動的に発見するため，まず隠語 DB から語を複数選び (図 2 中 (1))，それをクエリとして Web 検索を行う (図 2 中 (6))．ページのアンダーグラウンド系 BBS 判定には，ページのアンダーグラウンド性の判定と，ページが BBS かの判定が必要である．ページのアンダーグラウンド性判定では，文書のアンダーグラウンド性を表す指標として UG-score を導入し，検索結果の個々のページに対し UG-score を計算する (図 2 中 (9))．この値が閾値より大きいものを，アンダーグラウンドなページと判定する．ページの BBS 判定では，BBS に多い日付表現が発見できるか判定し，こうしてアンダーグラウンド系 BBS と判定されたページは，BBS-DB に加えられる (図 2 中 (10))．

続いて，発見した BBS の新着投稿を取得する (図 2 中 (4))．アンダーグラウンド系 BBS の投稿にはアンダーグラウンドな投稿以外もあるため，まず新着投稿の UG-score を算出する (図 2 中 (11))．UG-score が閾値以上ならアンダーグラウンドな投稿と判断し，投稿中の各語に対し，その語と周辺語をクエリとして Web 検索を行う (図 2 中 (7))．次に，検索結果と元の投稿をクラスタリングし，語法で検索結果を分類する．個々のクラスタで語を隠語として用いているか判定するため，元の投稿が属するクラスタに含まれるページの UG-score 値を算出し，その平均を求める．この値がクラスタのアンダーグラウンド性を表すとみなし，この値が閾値以上で，かつクラスタが BBS を含んでいれば，このクラスタは語を隠語として用いているとみなす．このとき，元の語を隠語と判定して隠語 DB に加える (図 2 中 (12),(13))．

以上の構成要素のうち，次の 3.2 節では，UG-score の計算に用いる SVM に対する学習の詳細を述べる．本研究では高精度な隠語判定を行うため，1 章での隠語の社会的機能にある仲間意識に着

目し，隠語を用いる社会集団が醸成する集団の言語的特徴を，既知の隠語に囚われず，より柔軟に認識し判断する．本研究では，こうした柔軟な判定を計算機で再現するため，SVM と潜在 Dirichlet 配分法 (LDA) を用い，これを隠語 DB とコーパスとなる BBS の投稿を用いて学習させる (図 2 中 (2),(5))．これらの手法を組み合わせ，隠語単位ではなく，隠語以外の語も多く含むトピック単位の機械学習を行うことで，社会集団の言語的特徴を学習できる．

続いて，3.3 節では，文書のアンダーグラウンド性を表す指標として UG-score を導入する．UG-score は，文書中に隠語 DB の語がどれだけ含まれるか (図 2 中 (3)) と，上述の SVM の出力値 (図 2 中 (8)) から計算される．文書の UG-score は，与えられた文書がどの程度「アンダーグラウンド的」かを表す数値であり，提案手法の随所で用いる重要な指標である．

以上の準備を基に，3.4 節でアンダーグラウンド系 BBS の発見手法を，3.5 節で発見した BBS からの新たな隠語の発見手法を述べる．最後に 3.6 節で，本研究で扱った薬物系や出会い系の隠語発見に特化した精度向上の手法を述べる．

3.2 SVM と LDA による学習

3.1 節で述べた通り，高精度な隠語判定には，隠語を用いる社会集団が醸成する集団の特徴を柔軟に認識し判断することが必要である．こうした柔軟な判断を計算機で再現するため，SVM による教師あり学習を行う．本研究では，薬物関連と出会い系関連の隠語検出を目的として，2 ちゃんねる「薬・違法板」[1] の過去ログ (投稿数 103,663) と pinkbbs「お水出会い系板」[2] の過去ログ (投稿数 223,698) のそれぞれで SVM を学習させた．

これらの BBS $\mathbb{B}$ を投稿 A_i の集合 $\mathbb{B} := \{A_i | i = 1, \cdots, n\}$[3]，BBS $\mathbb{B}$ に現れる語の集合を W とする．助詞などのストップワードは精度や速度の低下要因となるため，投稿から予め除く．また，今回対象とする薬物や出会い系のカテゴリでは，密売人などとの直接の接触を行う手段にメールが用いられる．密売人などの投稿ではメールアドレスが含まれやすいため，メールアドレスを特殊な隠語とみなす．その上で，投稿 A_i を W の部分集合とする．

命題 P に対し，δ_P を P が真のとき 1，偽のとき 0 と定義し，行列 D を $D := (\delta_{j \in A_i})$ と定義する．すなわち，D の (i, j)-成分は、投稿 A_i で語 j が現れるとき 1 になる．

次に，D に LDA を用いトピック抽出を行う．トピック単位でスコア付けすることで，SVM はより柔軟な判定を行える．SVM と LSA を組み合わせて精度を上げる研究には，Kwok[2]，Shima ら [5] などがある．

抽出されたトピック集合を $\mathcal{T} := \{T_1, \cdots, T_{N_T}\}$ とすると，T_l は語とその生起確率の組の集合 $T_l = \{(j, p_j) \in W \times [0, 1]\}$ として表される．このとき，$s_l := \sum_{j \in T_l} p_j \delta_{j:\text{隠語}}$ をトピック T_l の学習用スコアに用いる．そして，$S_A := \sum_l p(A|T_l) s_l$ を投稿 $A (\in \mathbb{B})$ の学習値として SVM に学習させる．

3.3 UG-score

続いて，文書のアンダーグラウンド性を表す指標として UG-score を導入する．まず，与えられた語がどの程度「隠語らしい」かを表す $[0, 1]$ 上の値で語の UG-score を定義する．

次に，既知の隠語 j とその語の UG-score w_j の組 (j, w_j) からなる辞書を D_{Jargon} とする．初期状態で D_{Jargon} には既知の隠語 DB の語 j を登録し，これらの語については $w_j := 1$ とする．

$$D_{\text{Jargon}} := \{(j, w_j) | j : \text{隠語}, w_j : j \text{ の UG-score}\}$$

語の UG-score を用い，文書 d の UG-score u_d を定義する．

$$u_d := \lambda \frac{\sum_{j \in d \cap D_{\text{Jargon}}} w_j}{\sum_{j \in D_{\text{Jargon}}} w_j} + (1 - \lambda)\theta_d \ (\lambda \in [0, 1]) \tag{1}$$

右辺の第 1 項は D_{Jargon} の元で文書 d にも含まれる隠語の割合，第 2 項の θ_d は 3.3 節で学習させた SVM に文書 d を入力したときの出力値である．SVM の出力値を用いることで，D_{Jargon} に含まれない未知の隠語のみを含む文書にも適切な UG-score が与えられる．また，経験的に決定した定数 λ でこれらの因子を重み付ける．文書集合 C の UG-score U_C を，C の各元の UG-score の平均として，$U_C := \frac{1}{|C|} \sum_{d \in C} u_d$ と定義する．

3.4 アンダーグラウンド系 BBS の発見

1 章で述べた通り，本研究では新たな隠語の発見元として BBS への投稿を用いる．そこで本節では，隠語を含む投稿がなされる BBS の発見手法を述べる．既知のアンダーグラウンド系 BBS b とその UG-score u_b の組 (b, u_b) の辞書 (BBS-DB) を D_{BBS} とするとき，この手法の概略は次のようになる．

まず，隠語 DB D_{Jargon} から隠語 j_1, j_2 を選び，「bbs j_1 j_2」という語句をクエリとして Web 検索を行う [4]．検索結果を Web ページの集合 $\{b_i\}$ としたとき，各ページ b_i がアンダーグラウンド系 BBS であるかどうかを判定するため，次の手順を行う．

1. b_i が既に BBS-DB D_{BBS} に含まれていれば，b_i は既知の BBS なので終了する．
2. b_i の UG-score u を算出する．
3. u が閾値 μ_{BBS} 以上か調べ，μ_{BBS} 以下なら b を負例 D^- に加えて終了する．μ_{BBS} 以上なら，b を正例 D^+ に加える．
4. b_i に日付表現が多数含まれることを調べることで b が BBS か調べ，BBS でなければ終了する．これは，BBS には投稿日時を表す日付表現が多数現れることに着目したもので，blog 記事を自動検出する南野ら [10] の手法を応用した．
5. b_i とその UG-score u を BBS-DB D_{BBS} に追加する．

すべてのページ b_i に対しアンダーグラウンド系 BBS か判定した後，式 2 で j_1, j_2 の UG-score を更新する [5]．ただし，α, β, γ は経験的に決定される定数である．これにより j_1, j_2 の UG-score は，発見できたアンダーグラウンドなページが多ければ大きく，少なければ小さくなる．その結果，次回以降に文書の UG-score を計算する際に，j_1, j_2 の重みが変化し，より良いアンダーグラウンド性判定が行えるようになる．

$$w_{j_k} := \alpha w_{j_k} + \beta U_{D^+} - \gamma(1 - U_{D^-}) \ (k = 1, 2) \tag{2}$$

3.5 新たな隠語の発見

続いて，3.4 節で得た BBS-DB D_{BBS} から新たな隠語を発見する手法を述べる．まず，D_{BBS} に登録された BBS $\mathbb{B}$ の新着投稿を取得する．新着投稿の取得には，それぞれの投稿が HTML 中のどの位置に存在するか判定する必要がある．南野ら [10] は，blog の本文を検出するために日付表現に隣接するタグに着目したが，本研究ではより簡易な手法として，BBS の投稿中で多用される改行に着目した．HTML では br タグが改行を表すため，BBS 中に含まれるすべての br タグの XPath を取得し，どの XPath で br タグが最も現れやすいか判定する．これにより，各投稿の本文の HTML 構造中での位置を検出できる．

[1] http://anago.2ch.net/ihou/

[2] http://kilauea.bbspink.com/pub/

[3] BBS によっては投稿をまとめたスレッドという単位が存在する場合があるが，スレッドは無視している．

[4] 本研究では，Google の Custom Search API (https://developers.google.com/custom-search/) を用いた．

[5] 式 2 は，適合フィードバックで用いる Rocchio の更新式 [4] を応用したものである．

次に，新着投稿 $A(\in \mathbb{B})$ の UG-score u_A を計算し，u_A が閾値 μ_{Article} 以下なら隠語が含まれる確率が低いと判断し，次の投稿へ移る．u_A が μ_{Article} 以上なら，この投稿 A に含まれるそれぞれの語 j について，次の手順で j の隠語判定を行う．

1. j が助詞などのストップワードであれば終了する．
2. j が既に隠語 DB D_{Jargon} の元ならば，j は既知の隠語なので終了する．
3. 橋本ら [7] の手法を用い，j と投稿中での j の周辺語 j', j'' をクエリとして Web 検索する．
4. 検索結果のページ集合 $\{p_i\}$ に元の投稿 A を加えた集合 $\{p_i\}\cup\{A\}$ から LDA によるトピック抽出を行う．ここで，抽出するトピック数は事前に指定しておく．次に，トピックをクラスタとみなしてクラスタリングを行い，各文書は生起確率の最も高いトピックのクラスタに属するとみなす．これにより生成されるクラスタは，j の用法別に分類される．
5. A を含むクラスタ C の UG-score U_C を計算する．
6. U_C が閾値 μ_{Cluster} 以下なら，クラスタ C はアンダーグラウンドなページ集合でないと判断し終了する．
7. U_C が μ_{Cluster} 以上，かつ C がある程度 BBS を含む場合，j は隠語だと判断し，組 (j, U_C) を D_{Jargon} に加える．1 章で述べた通り隠語はアンダーグラウンド系 BBS に偏在すると考えられるため，C に BBS がある程度含まれるか 3.4 節と同様にして確認する．

3.6 追加的なチェック

前節までの手法に加え，本研究では隠語発見の精度を上げるため，隠語 DB D_{Jargon} に語を追加する前に次の 2 つの追加的なチェックを行った．まず，本研究では新しい隠語を発見するので，C に含まれるページの最終更新日時がある程度新しいかを確認する．次に，政府機関や学術関連の Web ページで用いられる語は既によく知られた語だと考えられるので，C に.go.jp, .lg.jp, .ed.jp, .ad.jp などのドメインのページが含まれないか確認する．

4. 評価実験

第 3 章の提案手法に基づき，評価実験を行った．LDA と SVM の実装には Gensim[3] と LIBSVM[1] を用いた．また，本研究の手法は，既知の隠語を用いて新たな隠語を発見するスノーボールサンプリング型の手法なので，最初に既知の隠語の初期データベースが必要となる．今回の実験では，Jetrun テクノロジ株式会社 [6] の「隠語・誘導語データベース」[7] のうち,「薬物」「出会い」カテゴリの 500 語ずつを，別々にスノーボールとして用いた．

本研究では隠語 DB に「薬物」カテゴリと「出会い系」カテゴリの語を用いたので，評価実験もこれらそれぞれのカテゴリで行った．また，UG-score の定義式 1 中の定数 λ を $\lambda = 0.75$ とした．加えて，精度向上のため，文書中で隠語 DB の語が 3 語以下の場合，式 1 の第 1 項を 0 とした．

3 章冒頭で述べた通り，提案手法はアンダーグラウンド系 BBS の発見部と新たな隠語の発見部が並列に動くため，評価実験もこれらの各部に対し行った．前者の評価実験を 4.1 節で，後者の評価実験を 4.2 節で述べる．

[6] http://www.kkyg.jp/

[7]「誘導語」は Jetrun テクノロジ独自の呼称で，学校非公式サイトなどで用いられる，過激な発言を助長する語を指す．

表 1: クエリ「BBS ケタラール サルビア」での動作結果

Table 1. Results by using 「BBS ケタラール サルビア」 as a query

Web ページの題名	UG-score 値	UG-score 第 1 項	UG-score 第 2 項	判定
サルビア・ディビノラムの思い出 Salvia Divinorum	0.0812	0.0498	0.1755	アンダーグラウンド系 BBS
生きたまま人間の手足の皮膚を溶かす方法とは - 復讐掲示板	0.0493	0.0298	0.1077	アンダーグラウンド系 BBS
薬物で幽体離脱するスレ - livedoor したらば掲示板	0.0460	0.0079	0.1604	アンダーグラウンド系 BBS
ケミカルドラッグ掲示板ログ 200005	0.0570	0.0159	0.1802	アンダーグラウンド系 Web ページ
最近幽体離脱にはまった【避難所】 - livedoor したらば掲示板	0.0428	0.0	0.1715	アンダーグラウンド系 BBS
合法ハーブと合法ドラッグと合法ケミカル研究所	0.0347	0.0219	0.0734	非アンダーグラウンド系 Web ページ

4.1 アンダーグラウンド系 BBS の発見

3.4 節で述べたアンダーグラウンド系 BBS の発見手法が適切に動作するか調べるため，システムを 20 回動作させた際の検索結果ページがアンダーグラウンド系 BBS か人手で判別し，その結果とシステムの動作結果を比較した．実験では，3.4 節での定数 μ_{BBS} を，薬物系の場合は $\mu_{\text{BBS}} = 0.04$，出会い系の場合は $\mu_{\text{BBS}} = 0.02$ とした．また，UG-score の更新式 2 中の定数 α, β, γ を $\alpha = \beta = \gamma = 0.5$ とした．

以上の条件の下でシステムを動作させた結果のうち，上位 6 件を表 1 に示す．この動作では，システムは「BBS ケタラール サルビア」をクエリとして検索し，検索結果に含まれるページがアンダーグラウンド系 BBS か判定した．第 1 列は検索結果に含まれるページの題名，第 2 列がそのページの UG-score 値，第 3 列，第 4 列は式 1 中の，定数 λ による重み付け前の各項の値，第 5 列がシステムの各ページに対する判断である．

表 1 中のページはすべてアンダーグラウンド系 Web ページであり，上位 3 件と第 5 位はアンダーグラウンド系 BBS である [8]．第 6 位の「合法ハーブと合法ドラッグと合法ケミカル研究所」は合法ハーブについて扱っているページであり，アンダーグラウンド系 Web ページと判断されるべきであるが，UG-score が閾値 $\mu_{\text{BBS}}(= 0.04)$ 以下となったことから，誤って非アンダーグラウンド系ページと判断している．

同様の動作を繰り返した際の薬物系の BBS に関する実験結果を表 2 に，出会い系の BBS に関する実験結果を表 3 に示す．実験の結果，薬物系と出会い系のページのアンダーグラウンド系 BBS 判定では，平均 73%と平均 74%の適合率を得た．

表 2，表 3 から，文書のアンダーグラウンド性判定で良い結果が出るクエリ (表中で強調) と結果が芳しくないクエリの二種類があることが分かる．次の 2 点がこの原因として考えられる．

- 実装したシステムでは検索に用いる 2 語を完全にランダムに選んでいるため，いずれも薬物関連の隠語であっても，必ずしも関連する語が選ばれるとは限らない．例えば，表 2 中の「民剤」は睡眠導入剤の隠語,「ヤーケー」はコカインの隠語である．睡眠導入剤とコカインの常用者の重なりは小さいと考えられるが，今回の実験ではこのようなことを考慮していない．この問題の解消手法として，隠語 DB を隠語の意味でクラスタリングし，クエリとする隠語を同一のクラスターから選ぶことが考えられる．
- クエリに選ぶ語によってはアンダーグラウンドでない Web ページが多くなってしまう．例えば，表 2 中の「bush」は大麻の隠語だが，これをクエリに含めると，アメリカの Bush 元大統領に関するページが多くなり，大麻に関するページが減ってしまう．一方，麻酔薬の商標「ケタラール」のように

[8] 第 4 位の「ケミカルドラッグ掲示板ログ 200005」は BBS のログで静的な Web ページであることから，BBS でないと判断するのが妥当である．

表 2: アンダーグラウンド系 BBS の発見 (薬物系)

Table 2. Results of discovery of underground BBS (on drugs)

クエリ		アンダーグラウンドと判定			アンダーグラウンド系 BBS と判定		
j_1	j_2	総数	誤判定	適合率	総数	誤判定	適合率
ベイ中	ツイスト	9	9	0.0	1	1	0.0
売人	bush	3	3	0.0	1	1	0.0
カラス	ざらめ雪	10	9	0.1	1	1	0.0
マジック							
コーヒー	チョーク	7	7	0.0	2	2	0.0
赤ネタ	ピンクダイヤ	3	2	0.4	2	1	0.5
カラス	ホワイト	3	2	0.4	1	0	1.0
インディカ	ヤク中	**9**	**1**	**0.9**	**6**	**0**	**1.0**
雪	やせ薬	**5**	**2**	**0.6**	**1**	**0**	**1.0**
オクレ兄さん	サンペドロ	4	2	0.5	0	0	-
赤ちゃん	Shroom	1	1	0.0	0	0	-
ブラック							
ビューティー	ブラウンズ	1	1	1.0	0	0	0.0
マグルス	赤ちゃん	4	4	0.0	0	0	-
Mary Jane	SP	1	1	0.0	0	0	-
MDA	whack	1	1	0.0	0	0	-
エリミン	メアリージェイン	**12**	**3**	**0.75**	**0**	**0**	**-**
	ピンクアンド						
アキアジ	グリーンアンプ	2	2	0.0	0	0	-
ケタラール	サルビア	**13**	**1**	**0.93**	**6**	**0**	**1.0**
	マザーオブ						
ワンジー	パール	0	0	-	0	0	-
民剤	ヤーケー	6	0	1.0	1	0	1.0
joint	パールバティ	2	2	0.0	0	0	-
合計		96	53	0.45	22	6	0.73

表 3: アンダーグラウンド系 BBS の発見 (出会い系)

Table 3. Results of discovery of underground BBS (on prostitution)

クエリ		アンダーグラウンドと判定			アンダーグラウンド系 BBS と判定		
j_1	j_2	総数	誤判定	適合率	総数	誤判定	適合率
別荷	大人の出会い	6	1	**0.84**	1	0	1.0
143	ムービーモデル	4	4	0.0	0	0	-
プロフ	面接落ち	6	5	0.17	1	1	0.0
かまちょ	大人の出会い	**8**	**0**	**1.0**	**4**	**0**	**1.0**
TEL エッチ	直電	**4**	**0**	**1.0**	**0**	**0**	**-**
スペ	友募	0	0	-	0	0	-
CB	ホ別	**4**	**1**	**0.75**	**3**	**0**	**1.0**
リア工	場所ナシ	6	3	0.5	2	1	0.5
ドム	チャカレ	**10**	**2**	**0.8**	**3**	**1**	0.67
NP	派遣します	**3**	**0**	**1.0**	**3**	**0**	**1.0**
サポ希望	サシ	12	9	0.25	5	4	0.2
コチャ	フレ募集	1	1	0.0	0	0	-
ブッチ	地蔵	12	8	0.33	4	3	0.25
イエローデブ	申し	6	4	0.33	3	0	1.0
パパ	サポート	9	5	0.45	1	0	1.0
SF	円	2	2	0.0	0	0	-
番交換	イエローデブ	5	3	0.4	3	0	1.0
CC	チャ専	2	2	0.0	1	0	1.0
かわぼ	ケーバン	**11**	**1**	**0.91**	**4**	**0**	**1.0**
スタビ	援交際	**8**	**0**	**1.0**	**0**	**0**	**-**
合計		119	51	0.58	38	10	0.74

固有名詞に由来する語や，隠語以外の用法がない語をクエリとした場合は，比較的良い結果が得られる．

また，表 2，表 3 から，薬物系より出会い系の方がアンダーグラウンド性判定の誤報率が低いことが分かる．この原因としては，まず出会い系サイト自体は違法性がないため，犯罪に結びつきやすい薬物系の話題を扱うサイトより母数が多いことが考えられる．加えて，出会い系の場合に用いたクエリに比べ，薬物系の場合に用いたクエリは「カラス」「ツイスト」などの一般名詞が多いため，誤検出が多くなったと考えられる．

4.2 新たな隠語の発見

3.5 節で述べた隠語の発見手法を実装し，評価実験を行った．実験では，3.5 節での定数を $\mu_{\text{Article}} = 0.01, \mu_{\text{Cluster}} = 0.01$ とした．また，今回の実験では精度向上や高速化のため，次の 2 点を変更した．まず，すべての語 j に対し隠語判定を行うのではなく，Yahoo! JAPAN のキーフレーズ抽出 API[9] を用いて対象とする語を制限した．この API は日本語文を解析して特徴語を抽出し，0 以上 100 以下の整数値で重要度を算出できるが，実験では重要度 50 以上の語のみを処理対象とした．次に，3.5 節では UG-score が μ_{Cluster}

[9] http://developer.yahoo.co.jp/webapi/jlp/keyphrase/v1/extract.html

表 4: クエリ「ドラ アンフェタミン」の検索結果

Table 4. Results of the query 「ドラ アンフェタミン」

	ページの題名
1	イリーガル系ドラッグの解説 ブリトニー・スピアーズの元マネージャーがドラッグ疑惑を暴露 ...
2	覚醒剤 - Wikipedia メチレンジオキシメタンフェタミン - Wikipedia ドラッグ用語集 医薬品一覧 - Wikipedia 合法ハーブ研究所～合法ドラッグの歴史～ ドラッグとは - はてなキーワード
3	パラクロロアンフェタミンの英語・英訳 - 英和辞典・和英辞典 Weblio 辞書 モーマス，フィル・ウィルソン，ビル・ドラモンド，ベイビー・アンフェタミン ... 覚醒剤は素晴らしい
4	アンフェタミンによりゾンビ化してしまったような，瞳孔を見開く中毒女性の ...
5	覚せい剤は実は非常に安全な薬です（研究報告） 覚醒剤アンフェタミンの原材料，8 割が中国からの密輸品―米国 (Record ... 釣られたお？ 【Drugs-forum より】(メス) アンフェタミン等使用後の回復法 ... 『ボイド・イズ・マイ・アンフェタミン』 笹川 作 エクスタシー，XTC と呼ばれる MDMA - STOP the DRUG ドラッグ Cannabis Study House - ドラッグ・テスト ― ドラッグ・テストの種類と問題点 覚せい剤は実は非常に安全な薬です（研究報告） (「ドラ」を含む元の投稿)

表 5: クラスター 5 の UG-score 算出結果

Table 5. Results of computation of UG-score

ページの題名	UG-score 値	第 1 項	第 2 項
覚せい剤は実は非常に安全な薬です（研究報告）	0.02175	0.01197	0.05107
覚醒剤アンフェタミンの原材料，8 割が中国からの密輸品―米国 (Record ...	0.00204	0.0	0.00818
釣られたお？ 【Drugs-forum より】(メス) アンフェタミン等使用後の回復法 ...	0.01784	0.00998	0.04144
『ボイド・イズ・マイ・アンフェタミン』 笹川 作	0.03659	0.03393	0.04457
エクスタシー，XTC と呼ばれる MDMA - STOP the DRUG	0.01054	0.0	0.04218
ドラッグ	0.01371	0.0	0.05484
Cannabis Study House - ドラッグ・テスト ― ドラッグ・テストの種類と問題点	0.01204	0.0	0.04819
覚せい剤は実は非常に安全な薬です（研究報告）	0.02084	0.01197	0.04747
(「ドラ」を含む元の投稿)	0.01134	0.0	0.04538
合計	0.17138		
平均 (クラスター 5 の UG-score)	0.01904		

以上かの判定後にクラスタが BBS を含むか判定し，次に 3.6 節の追加的な判定を行うが，UG-score の計算が低速なため，BBS の判定と追加的な判定の後に UG-score の判定を行うようにした．

以上の条件下で,「ドラ」というドラッグを表す隠語を発見した際のシステムの動作を次に述べる．語「ドラ」を含む投稿では「アンフェタミン」という周辺語があったため，システムは「ドラ アンフェタミン」というクエリで Web 検索を行った．

検索結果として得られたページをクラスタリングした結果が表 4 である．表中での罫線はクラスターの境界を表し，便宜的にクラスターに番号を付している．

表 4 のクラスタリング結果を見ると，クラスター 2 は主に薬物に関する科学的なページからなる．また，クラスター 1,3,4 には薬物関連のニュースなどからなる．一方，元の記事を含むクラスター 5 には，ニュースサイトなども含まれているが,「覚せい剤は実は非常に安全な薬です（研究報告）」という 2 ちゃんねるのスレッドや薬物に関するフォーラムへの投稿など，アンダーグラウンドな内容を含むページが多いことが分かる．

次にシステムは，クラスター 5 に BBS が含まれるかを調べる．今回は 2 ちゃんねるのスレッドがクラスター 5 に含まれているので，これを BBS だと判定し，処理を続行した．クラスター 5 の各ページの UG-score を算出した結果を表 5 に示す．なお，表 5 の第 3 列，第 4 列については表 1 と同様に，重み付け前の UG-score の各項の値を表す．

表 5 よりクラスター 5 の UG-score は 0.01904 となり，μ_{Cluster}(= 0.01) を上回る．従ってクラスター 5 はアンダーグラウンドなページからなるクラスタとなり,「ドラ」は隠語と判定された．

同様にシステムを動作させたときに得られる結果の一部を，表 6 と表 7 に示す．表 6 は薬物系の，表 7 は出会い系の隠語を発見するシステムの動作結果である．なお，表中で強調した行は，隠

表 6: 新たな隠語の発見 (薬物系)

Table 6. Results of discovery of new jargons (on drugs)

判定対象の語	判定	隠語でない場合の理由
覚せい剤	×	クラスタに BBS がない
アンフェタミン	×	クラスタに BBS がない
ドラ	○	
鶴見済	○	
人格改造マニュアル	○	
ドラちゃん	○	
ベーハ	×	クラスタの UG-score が μ_{Cluster} 以下
半減期	×	クラスタに.go.jp ドメインのページが存在
メタンフェタミン	×	クラスタに BBS がない
ドパミン	×	クラスタに BBS がない
KDDI-TS3N UP.Browser	○	
耳かき	×	クラスタに BBS がない
耳かき 1 杯	×	クラスタに BBS がない
2 ちゃんねる	○	
ガンギマリ	○	

表 7: 新たな隠語の発見 (出会い系)

Table 7. Results of discovery of new jargons (on prostitution)

判定対象の語	判定	隠語でない場合の理由
風俗	×	クラスタに BBS がない
時点	×	クラスタに BBS がない
風俗嬢	○	
円光女	○	
マクソ	○	
ハピメ	○	
ポイント	○	
勝ち組	×	クラスタに BBS がない
中国人	×	クラスタに BBS がない
外人	○	
日本語	○	クラスタに.go.jp ドメインのページが存在
一見	×	クラスタの UG-score が μ_{Cluster} 以下
美人	○	
性病	×	クラスタに BBS がない
マグロ	×	クラスタに BBS がない

語を隠語と正しく検出できたと考えられる行である．いずれの実験でも，候補語中で 37.5%の適合率と 75%の再現率を得た．

表 6 で，最初の「覚せい剤」はクラスタに BBS が含まれないため隠語でないと判定された．これは，対象となった投稿を含むクラスタに，元の投稿以外に BBS が含まれなかったということを表す.「覚せい剤」という語は薬物の名称そのものであり，アンダーグラウンド系 BBS ではこうした直接的な表現を用いることは少ないため，この判定は適切と考えられる．

表 6 で 4 行目の「鶴見済」は人名，5 行目の「人格改造マニュアル」はその著作であるが，これらは隠語だと判定された．これらは薬物に関する隠語とはいえないので，これは誤判定である．

表 6 で 3 行目の「ドラ」，6 行目「ドラちゃん」はドラッグを，15 行目の「ガンギマリ」は「ガンガンに」薬物が「キマって」(効いて) いる状態を表す．また，表 7 で 4 行目の「円光女」は援助交際を行う女性を，5 行目「マクソ」，6 行目「ハピメ」は出会い系サイトの PCMAX とハッピーメールを表す．隠語 DB に含まれない未知の隠語であるこれらの語を発見できていることから，提案手法は有用性があるといえる．

5. おわりに

本研究では，新たな隠語の発見手法として，アンダーグラウンド系 BBS と隠語の発見を並列に行う手法を提案し，評価実験を行った．アンダーグラウンド系 BBS の発見手法では，UG-score の閾値を調整することで，比較的高い精度でアンダーグラウンド系 BBS を発見できた．また，新たな隠語の発見手法は，検出精度はあまり高くないが，隠語 DB にない隠語を発見できた．

研究課題としては，アンダーグラウンド系 BBS を発見する段階で，クエリとする隠語をランダムに選んだため，語の組み合わせによって良い結果が得られないことがある．この課題の解消手法としては,「スピード」と「アイス」のように意味の近い隠語の組でクエリを構成することが考えられる．また，提案手法は長期間の運用を想定し，時間の経過による隠語やアンダーグラウンド系 BBS の盛衰も考慮している．今回の評価実験では，そうした長期的な運用による隠語検出精度の変化について十分な検証をしていない．しかしながら，隠語辞書に登録された語の UG-score を低くし，アンダーグラウンド系 BBS の発見への貢献により UG-score を増加させるしくみを導入していることもあり，長期的な隠語発見の精度低下はさほど大きくないと考えられる．

【謝辞】

本研究では，Jetrun テクノロジ株式会社から購入した隠語データを利用させていただきました．長期に亙る契約交渉を辛抱強くご担当いただいた，Jetrun テクノロジ株式会社の佐久川様と小橋川様に深く感謝申し上げます．

【文献】

[1] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. ACM TIST, Vol. 2, pp. 27:1–27:27, 2011. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[2] James Tin-Yau Kwok. Automated text categorization using support vector machine. In In Proc. of ICONIP, pp. 347–351, 1998.

[3] Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In Proc. of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[4] J. J. Rocchio. Relevance feedback in information retrieval. Information Storage and Retrieval: Scientific Report ISR-9, 1965.

[5] K. Shima, M. Todoriki, and A. Suzuki. SVM-based feature selection of latent semantic features. Pattern Recogn. Lett., Vol. 25, No. 9, pp. 1051–1057, 2004.

[6] 山本英子, 梅村恭司. 辞書を用いない関連語リストの構築方法. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2002, No. 20, pp. 81–88, 2002.

[7] 橋本広美, 木下嵩基, 原田実. フィルタリングのための隠語の有害語意検出機能の意味解析システム sage への組み込み. 情報処理学会研究報告. SLP, 音声言語情報処理, Vol. 2010, pp. 1–6, 2010.

[8] 那小川. 潜在的意味解析による中国語のインターネット新語に関する研究, 2012. 修士論文, 東京大学大学院 情報理工学系研究科.

[9] 松葉達明, 里見尚宏, 桝井文人, 河合敦夫, 井須尚紀. 学校非公式サイトにおける有害情報検出. 信学技報, Vol. 109, pp. 93–98, 2009.

[10] 南野朋之, 鈴木泰裕, 藤木稔明, 奥村学. blog の自動収集と監視. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2004, No. 23, pp. 129–136, 2004.

[11] 米川明彦. 集団語の研究 (上巻). 東京堂出版, 2009.

大西 洋　Hiroshi OHNISHI

京都大学大学院情報学研究科修士課程在学中.

田島 敬史　Keishi TAJIMA

京都大学国際高等教育院教授，情報学研究科併任．1996 年東京大学大学院理学系研究科情報科学専攻博士課程修了，博士（理学）．データベースシステム，情報検索，Web からの情報抽出に関する研究に従事．日本データベース学会会員.