DEIM Forum 2013 C10-6

語の出現の偏りに基づく新たな隠語の発見

大西　洋† 田島　敬史††

† 京都大学工学部情報学科　〒606–8501 京都府京都市左京区吉田本町 36–1
†† 京都大学大学院情報学研究科　〒606–8501 京都府京都市左京区吉田本町 36–1
E-mail: †ohnishi@dl.kuis.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　有害情報や違法行為を隠蔽しつつ情報伝達を行うために，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出現することを利用して，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って，語が隠語かを判定する．また，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掲示板の発見を並行して行う．

キーワード　情報抽出，隠語，Web マイニング，情報フィルタリング

1. はじめに

1.1 背　景

隠語とは,「特定の職業や身分に属する限られた人々の間で，主として秘密を守ったり，あからさまにいうのを避けたりするために用いる特別のことば」[13] のことであり,「かくしことば」「符牒」などとも呼ばれる．例えば，薬物の呼称で「アイス」といえば覚醒剤を表し，出会い系の掲示板 (BBS) で「サポ」といえば援助交際を指す．ここでいう「アイス」「サポ」のような語が，本研究で扱う隠語である．言語学上は，こうした隠語は集団語という語集合の一部に属する．

集団語という学術用語を初めて用いたのは柴田 [16] である．この中で柴田は，隠語や職業語，スラングがいずれも集団の中で人工的に作られるものであることから，これらを包括して集団語と呼ぶことを提唱した．

渡辺 [19] は柴田の研究を発展させ，集団語を「社会集団が使用している，その集団に特有な，ないしは特徴的なことば」と定義し，集団語を社会学的観点で考察した．また，集団語を隠語か否かで分類し，非隠語が職業語やスラングを含むとした．

これに対し米川 [18] は，集団語が社会集団を基盤として成立することから，隠語か非隠語かで集団語を分類するのは適切でないとして，渡辺による集団語の分類を批判した．米川は，集団語を「特定の機能的社会集団 (血縁的・地縁的でない) に特有な，あるいは特徴的な仲間内の通用語」と定義した (注1)．米川はここでいう「機能的社会集団」として，反社会的集団，職業的集団，被拘束集団，学生集団，趣味娯楽集団を挙げ，これらの社会集団で集団語を分類することが妥当とした．その上で，隠語を「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」と定義した．

米川 [18] は，隠語の社会的機能に次の 4 点を挙げており，これらの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**

1.2 現状の問題点

隠語の機能の一番目に挙げた所属集団の秘密を保持する機能は，集団の外にいる他者がその語を見ても内容が分からないようにして，情報の隠蔽を図るものである．特に，隠語を使用する社会集団が反社会的集団の場合においては，ここでいう「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アンダーグラウンドでの取引においては，露骨に違法行為を表す語を用いるのではなくこうした隠語を用いることで，治安組織による捜査の目を逃れることができる．また，治安組織で隠語が広く知られてしまうと隠語はその意味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，違法薬物取引や名誉毀損などの犯罪行為，あるいは援助交際の勧誘や自殺教唆などの有害情報の伝達において，Web を介在する事例が増加している．こうした事例の多くでは，情報のやりとりの足がかりにアンダーグラウンドな Web ページが利用される．特にアンダーグラウンド系 BBS では，業者や個人による薬物関連の有害情報の書き込みが多く見られる．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，Web を介在する犯罪が増加しうる．従って，新たに生まれた隠語の迅速な発見手法が喫緊の課題として求められている．

1.3 研究目的

そこで本研究では，前節で述べた現状を鑑み，アンダーグラウンド系 BBS で新たに生まれた隠語をできるだけ早期に発見することを目的とする．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を軽減し，犯罪抑止や有害

(注1)：学術用語も集団語と類似しているが，学術用語は社会集団外でも正式に使用されることが想定されている．そのため原則として，学術用語は米川の定義では集団語から外されている．

情報の効率的なフィルタリングが可能になる．

1.4 本研究のアプローチ

前節の目的を達成するため，本研究では，新たな隠語の発見元として BBS の投稿に着目する．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，彼らにとっての危険性も大きい．そのため，秘密保持のために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，一定以上の閲覧者がいるページであるともいえる．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．

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 の投稿中の各語とその周辺にある語をクエリとして Web 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，語を隠語と判定する．

秘密保持のために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，短期間のうちに使われなくなっていく．かつて 2 ちゃんねるの薬・違法板 [1] は薬物関係の談義が盛んであったが，治安組織の監視が入り現在は寂れている．新たな隠語を迅速に発見するには，最新のアンダーグラウンド系 BBS も発見していく必要がある．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，その検索結果として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．

上述のように本研究では，最新の隠語の発見に最新のアンダーグラウンド系 BBS を利用し，最新のアンダーグラウンド系 BBS の発見に最新の隠語を利用する．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，両者を並行して漸次更新することで，興廃の激しい隠語やアンダーグラウンド系 BBS を恒常的に発見することを可能とする．

1.5 本論文の構成

次章以降の本論文の構成は，概ね次のようになっている．

第 2. 章では，隠語発見や情報フィルタリングに関する先行研究を説明し，本研究との違いを述べる．第 3. 章では，新たな隠語を発見するための提案手法を述べる．第 4. 章では，提案手法を実装し動作させることで，提案手法の有効性を検証する．最後に第 5. 章で，本研究で得た知見と今後の研究課題を述べる．

2. 関 連 研 究

本研究で発見する新たな隠語は新語の一種とみなすことも可能である．新語発見の研究としては，那 [14] による潜在意味解析 (LSA) を用いた Web からの中国語の新語発見の研究などがある．那は，既存の新語リストと収集した Web ページに LSA を適用した結果の類似度を比較することで，リスト中の新語と類似度が高い語を新語と判定している．そのためこの手法では，既存の新語リストにない概念を表す新語の取得は難しい．

また，隠語発見は同義語や関連語を発見する研究とも関わりがある．山本ら [11] は，未知語を含む Web ページから，未知語のうち前後の文字列が類似するものを同義語と判定している．この手法は辞書が不要という利点があるが，未知語を含む Web ページを探す手法は考察されていない．

これらの先行研究の研究対象は一般の Web ページを対象としており，特定の領域の Web ページに限定されたものではない．しかし，本研究では単に新語や関連語を発見するのみならず，それが隠語であることを判定する必要がある．そのため，アンダーグラウンド系 BBS のみに偏在する新語を発見することが重要であり，対象をアンダーグラウンド系 BBS に限定している．アンダーグラウンド系 BBS は一般の Web ページに比べて存続期間が短く，また発見が困難なものが多い．従って，本研究では先行研究と異なり，アンダーグラウンド系 BBS を発見する手法が必要となる．

情報フィルタリングの観点からも多くの先行研究がある．松葉ら [15] は学校非公式サイトでの有害情報のやりとりを規制するため，サポートベクタマシン (SVM) を用いたフィルタリングを行った．この手法では BBS の投稿を SVM の入力とし，有害語を含む BBS の投稿を有害と学習させ，投稿のフィルタリングを行う．ただ，有害語は手作業で収集するため，未知の有害語を自動収集する手法を研究課題としている．

橋本ら [12] は，与えられた語の周辺にある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．ただこの研究では，例えば「草」という隠語を検出するためには，「指定駅で草を手渡します」のように，語が隠語として使われている文例を予め人手で集めておく必要がある．そのため，そうした文例が無い状態では隠語かどうかの判定を行うことができない．本研究では，周辺語を用いるという着眼を活かしつつ，より自動化された隠語判定の手法を提案する．

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隠語を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．

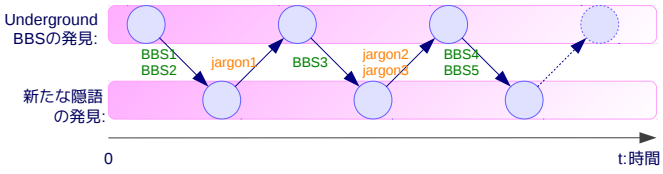

図 1 提案手法の概念図

1.4 節で述べたように，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 を発見する部分と新たな隠語を発見する部分が存在し，これらが並列に動作する．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) に追加する．次に BBS-DB の BBS より取得した新着投稿から，隠語「jargon1」を発見し，隠語データベース (隠語 DB) に追加する．続いて，隠語 DB より取得した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，隠語の盛衰やアンダーグラウンド系 BBS の興廃への柔軟な対応が可能となる．

より詳細な提案手法の全体図を図 2 に示す．図中の四角は提案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．青色の構成要素は本研究で実装した部分を，橙色の構成要素は提案手法で利用もしくは作成した辞書を，桃色の構成要素は提案手法で収集もしくは利用した Web 上の資源を表す．また図中で青色の矢印は提案手法での処理の流れを，橙色の矢印は辞書を起点とする辞書データの入力もしくは構成要素を起点とする辞書データの修正を，桃色の矢印は資源からのデータの入力を表す．

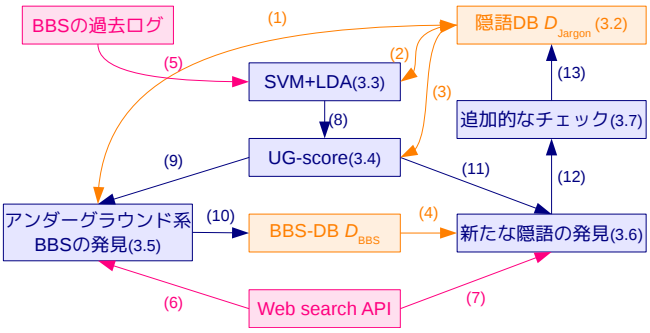

図 2　提案手法の全体図

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．

1.4 節で述べたように，本研究では，隠語発見のために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．まず，ページのアンダーグラウンド性判定のため，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．検索結果の個々のページについて UG-score を計算し (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．次に，ページが BBS か判定するため，BBS に多い日付表現が発見できるかを判定する．以上のように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 が閾値以上の投稿は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，語法別に検索結果を分類する．個々のクラスタで語が隠語として用いられ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．この値はクラスタのアンダーグラウンド性を表していると考えられるので，この値が閾値以上で，かつクラスタに BBS が含まれていれば，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，語を隠語 DB に追加する (図 2 中 (12),(13))．

このようなシステムを実装するため，3.2 節から 3.4 節で，提案手法で用いた辞書や指標について述べる．

3.2 節では，本研究で使用した隠語 DB について述べる．1.1 節で述べた隠語の社会的機能の一つに秘密保持があるが，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，薬物摂取に用いられる道具や取引のための語など，薬物に関連する語も隠語化する必要がある．薬物の名前のみを隠語化しても，注射器や販売といった語をそのまま用いてしまうと，こうした語から薬物に関する話題であることが容易に推測されてしまう．そこで，注射器を「ジェクター」，販売を「手押し」と隠語で表現することで，他者が文意を推測することが困難になる．従って本研究では，隠語は他の隠語と共起しやすいという仮定をおき，既知の隠語 DB を用いて隠語判定を行う．

3.3 節では，UG-score の計算に用いる SVM に対する学習の詳細について述べる．本研究ではより高精度な隠語判定を行うため，先の仮定に加えて 1.1 節で述べた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手法として，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て，隠語単位で学習を行うのではなく，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．

3.4 節で，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ているか (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られる重要な指標である．

以上の準備を基に，3.5 節でアンダーグラウンド系 BBS を発見する手法を，3.6 節で発見した BBS から新たな隠語を発見する手法を述べる．最後に 3.7 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　隠語・誘導語データベース

本研究では，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を用いた．今回用いた隠語 DB は，Jetrun テクノロジ株式会社が提供している「隠語・誘

導語データベース」の一部 [4] である (注2)．この隠語 DB に含まれる語の一部を表 1 に示す (注3)．

隠語	カテゴリ	解説
円光	出会い系	援助交際の意味．
神待ち	出会い系	食事や金銭を提供する人を「神」と例え，待って（募集して）いる．
スピード	薬物	覚せい剤の一種,「メタンフェタミン」の隠語．覚せい剤の隠語で主に利用されている．
キメセク	薬物	薬物を使用しながら，性交渉を行うこと．

表 1　隠語・誘導語データベースに含まれるデータ例

隠語 DB では，隠語を「犯罪」「薬物」「自殺」「いじめ」「出会い」「その他」の 6 種類のカテゴリに分類しているが，本研究ではこのうち「薬物」「出会い」カテゴリの語を 500 語ずつ用い，これらを別々にスノーボールとして用いた．

3.3 SVM と LDA による学習

3.1 節で述べたように，高精度な隠語判定のためには，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，主に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」[6] の過去ログ (投稿数 223,698) をそれぞれ SVM の学習データとした．

これらの BBS $\mathbb{B}$ を投稿 A_i の集合 $\mathbb{B} := \{A_i | i = 1, \cdots, n\}$ (注4)，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，投稿から予め除いておく．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ための手段として，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ることが多いため，メールアドレスは特殊な隠語「EMAILADDRESS」に予め変換しておく．以上の処理を施した上で，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 として定義し，行列 D を $D := (\delta_{j \in A_i})$ と定義する．ここで，D は BBS $\mathbb{B}$ の tf(term-frequency) 行列を表している．

次に，D に対して LDA を用いてトピック抽出を行う．トピックを単位として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向上する研究として，Kwok [5]，Shima ら [9] などがある．

抽出されたトピック集合を $\mathcal{T} := \{T_1, \cdots, T_{N_T}\}$ とするとき，T_l は語と語の生起確率の組の集合として $T_l = \{(j, p_j) \in W \times [0,1]\}$ と表される．このとき，$s_l := \sum_{j \in T_l} p_j \delta_{j:\text{隠語}}$ とおき，s_l をトピック T_l の学習用スコアとして用いる．

与えられた投稿 $A (\in \mathbb{B})$ に対し $S_A := \sum_l p(A|T_l) s_l$ とおき，これを投稿 A の学習値として SVM に学習させる．

3.4 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 $[0,1]$ 上の実数値として語の 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．

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

初期状態で D_{Jargon} に登録されている語は隠語 DB に含まれる語 j で，これらの語については $w_j := 1$ とする．

語の UG-score に基づいて，文書 d の UG-score u_d を次のように定義する．

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

右辺の第 1 項は D_{Jargon} の元のうち文書 d に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を表している．SVM の出力値を活用することで，D_{Jargon} に含まれない未知の隠語のみを含む文書に対しても，適切な UG-score が与えられるようになる．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する．

続いて，文書の集合 C の UG-score を，C の各文書の UG-score の平均と定義する．

$$U_C := \frac{1}{\#C} \sum_{d \in C} u_d$$

3.5 アンダーグラウンド系 BBS の発見

1.4 節で述べたように，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 (注5)．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 であることを判定するため，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 であるので終了する．

（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 [17] の手法を応用した．

(注2)：「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．

(注3)：本論文中の隠語データの著作権は Jetrun テクノロジ株式会社が保持しており，転載・再利用は禁止されている．本論文では Jetrun テクノロジ株式会社より特別に許諾を得て，データの一部を公開している．

(注4)：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．

(注5)：本研究では，Google が提供する Custom Search API [3] を利用した．

（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 に対してアンダーグラウンド系 BBS の判定を行った後，式 2 で j_1, j_2 の UG-score を更新する [注6]．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．

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

以上の手順はアルゴリズム 1 のように実装される．

Algorithm 1 アンダーグラウンド系 BBS の発見手法

```
D+ = D− := {}
j1, j2 := ChooseFrom(D_Jargon)
for b in WebSearch("bbs j1 j2") ∩ ¬D_BBS do
    u := UG-score(b)
    if u < μ_BBS then
        D− += b
        continue
    end if
    D+ += b
    if ! isBBS(b) then
        continue
    end if
    D_BBS += (b, u)
end for
UpdateScore(j1, j2, D+, D−, α, β, γ)
```

3.6 新たな隠語の発見

続いて，3.5 節で得た BBS-DB D_{BBS} を用いて新たな隠語を発見する手法について述べる．まず，D_{BBS} に登録されている BBS $\mathbb{B}$ の新着投稿を取得する．新着投稿の取得の際には，それぞれの投稿が HTML 中のどの位置に存在するかを判定する必要がある．南野ら [17] の手法では，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ているが，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 タグでは br タグが改行を表しているため，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を判定する．これにより，各投稿の本文の HTML 構造中での位置を検出することが可能になる．

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} に含まれていれば，j は既知の隠語であるので終了する．

（3） 橋本ら [12] 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} 以上で，かつ C がある程度 BBS を含む場合，j は隠語であると判断し，組 (j, U_C) を D_{Jargon} に加える．1.3 節で述べたように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 がある程度含まれているかを 3.5 節と同様の手法で確認する．

以上の手順はアルゴリズム 2 のように実装される．

Algorithm 2 新たな隠語の発見手法

```
for A in GetLatest(ChooseFrom(D_BBS)) do
    if UG-score(A) < μ_Article then
        continue
    end if
    for j in A ∩ ¬D_Jargon do
        if isStopWord(j) then
            continue
        end if
        for C in Clustering(WebSearch("j j' j''").Add(A)) do
            if A not in C then
                continue
            end if
            U_C := UG-score(C)
            if U_C > μ_Cluster and hasBBS(C) then
                D_Jargon += (j, U_C)
                break
            end if
        end for
    end for
end for
```

3.7 追加的なチェック

前節までの提案手法に加え，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追加的なチェックを行った．次に挙げる各項目は，アンダーグラウンドなページに関する経験的な知見に基づくものや，本研究で扱った薬物や出会い系の隠語発見に特化したものなどである．

- 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．
- 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であると考えられるので，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．

(注6)：式 2 は，適合フィードバックで用いられる Rocchio の更新式 [8] を応用したものである．

4. 評価実験

第 3. 章で提案した手法に基づき，Python でシステムを実装して評価実験を行った．なお，LDA の実装には Gensim [7] を，SVM の実装には LIBSVM [2] を用いた．

本研究では隠語 DB として「薬物」カテゴリと「出会い系」カテゴリの語を使用したので，評価実験もこれらそれぞれのカテゴリについて行った．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 に含まれる語が 3 語以下の場合，式 1 の第 1 項を 0 とした．

3. 章冒頭で述べたように，本研究での提案手法はアンダーグラウンド系 BBS を発見する部分と新たな隠語を発見する部分が並列に動作する．そのため，評価実験もこれらの各部に対して行った．4.1 節ではアンダーグラウンド系 BBS を発見する手法に対する評価実験を，4.2 節では新たな隠語を発見する手法に対する評価実験を述べる．

4.1 アンダーグラウンド系 BBS の発見

3.5 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 かを人手で判別し，その結果とシステムの動作結果を比較した．実験では，アルゴリズム 1 中の定数 μ_{BBS} を，薬物系の場合は $\mu_{\mathrm{BBS}} = 0.04$，出会い系の場合は $\mu_{\mathrm{BBS}} = 0.02$ 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 2 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．

Web ページの題名	UG-score 値	第 1 項	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

表 2　クエリ「BBS ケタラール サルビア」での動作結果

表 2 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 (注7)．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 $\mu_{\mathrm{BBS}}(= 0.04)$ 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 3 に，出会い系の BBS に関する実験結果を表 4 に示す．実験の結果，薬物系のページのアンダーグラウンド系 BBS 判定では平均 73%の適合率を，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 74%の適合率を得た．

クエリ j_1	クエリ j_2	アンダーグラウンド と判定 総数	誤判定	適合率	アンダーグラウンド系 BBS と判定 総数	誤判定	適合率
ペイ中	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 ピンクアンド	**12**	3	**0.75**	0	0	-
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 マザーオブ	**13**	1	**0.93**	**6**	0	**1.0**
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	パールパティ	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 (薬物系)

クエリ j_1	クエリ j_2	アンダーグラウンド と判定 総数	誤判定	適合率	アンダーグラウンド系 BBS と判定 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**

表 4　アンダーグラウンド系 BBS の発見 (出会い系)

表 3，表 4 を見ると，文書のアンダーグラウンド性判定では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．このような現象が見られる原因として，次の 2 点が考えられる．

- 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るといっても，必ずしも関連のある語が選ばれるとは限らない．例えば，表 3 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．睡眠導入剤を薬物として常用する層とコカインを常用する層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が考慮できていない．この問題を解消する手法として，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．
- クエリとして選ぶ語によってはアンダーグラウンドでないジャンルに属する Web ページが多くなってしまう．例えば，表 3 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

(注7)：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．

ル」のように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 3，表 4 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られたクエリに比べ，薬物系の場合に用いられたクエリは「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

3.6 節で述べた隠語の発見手法 (アルゴリズム 2) を実装し，システムが適切に動作するか実験した．今回の実験では精度向上や高速化のため，アルゴリズム 2 から次の点を変更した．

- アルゴリズム 2 の定数 $\mu_{\mathrm{Article}}, \mu_{\mathrm{Cluster}}$ を $\mu_{\mathrm{Article}} = 0.01, \mu_{\mathrm{Cluster}} = 0.01$ とした．
- アルゴリズム 2 中の $A \cap \overline{D_{\mathrm{Jargon}}}$ の要素 j に対するループ処理を行う際に，すべての j に対して処理を行うのではなく，Yahoo! JAPAN が提供しているキーフレーズ抽出 API [10] 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．
- アルゴリズム 2 中では UG-score が μ_{Cluster} 以上かの判定後にクラスタが BBS を含むか判定し，次に 3.7 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 5 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．

	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　クエリ「ドラ アンフェタミン」の検索結果

表 5 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 6 に示す．なお，表 6 の第 3 列，第 4 列については表 2 と同様に，重み付け前の UG-score の各項の値を表す．

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		

表 6　クラスター 5 の UG-score 算出結果

表 6 よりクラスター 5 の UG-score は 0.01904 となり，$\mu_{\mathrm{Cluster}}(= 0.01)$ 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 7 と表 8 に示す．表 7 は薬物系の隠語を発見するシステムの動作結果，表 8 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．

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 (薬物系)

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

表 8　新たな隠語の発見 (出会い系)

表 7 で，最初の「覚せい剤」はクラスタに BBS が含まれないことを理由に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 7 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，これらの固有名詞は隠語であると判定された．薬物に関する隠語を発見するというシステムの動作からいえば，これらは誤判定といえる．

表 7 で 3 行目の「ドラ」，6 行目「ドラちゃん」はドラッグを指し，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 8 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」はそれぞれ出会い系 Web サービスの PCMAX とハッピーメールを表す．これらはいずれも，隠語 DB に含まれていない未知の隠語である．これらの語が発見できていることから，本研究での提案手法は一定程度有用であるといえる．

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 の発見と隠語の発見を並列に行う手法を提案し，これを Python 及び Gensim [7]，LIBSVM [2] により実装した．

本研究で行った評価実験から，次の知見が得られた．

- 既知の隠語を辞書に用いてアンダーグラウンド系 BBS を発見する提案手法は，UG-score の閾値を薬物系で 0.04，出会い系で 0.02 程度にすることで，比較的高い精度でアンダーグラウンド系 BBS を発見することが可能である．
- アンダーグラウンド系 BBS の新着投稿から新たな隠語を発見する提案手法は，検出精度はあまり高くないが，隠語 DB に登録されていない隠語を発見することが可能である．

以上から，本研究で提案した手法は新たな隠語の発見に対して有用であるといえる．

本研究における研究課題としては，次のものが考えられる．

- アンダーグラウンド系 BBS を発見する段階で，現段階ではクエリとして用いる隠語はランダムに選んでいる．こうすることで多様なアンダーグラウンド系 BBS の発見が可能となるが，表 3 にあるように，クエリとする隠語の組み合わせによっては良い結果が得られないことがある．この課題を解消する手法として，クエリに用いる隠語の組を,「スピード」と「アイス」のように意味の近いもので構成することが考えられる．
- 本研究で実装したシステムは比較的長期間の運用を想定し，時間の経過による隠語の変化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，Web 検索 API の使用回数制限もあり，そうした長期的な運用によるシステムの隠語検出精度の変化について十分な検証ができていない．しかしながら，一度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 の発見への貢献をもって 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の影響はさほど大きくないと考えられる．
- 提案手法では Web 検索 API を用いるが，現在利用可能な Web 検索 API サービスのうち，本研究のようにデータの再利用を目的とした用途で利用可能なサービスは少ない．また，再利用目的で利用可能なサービスについても，1 日あたりの API 利用可能回数があるため，提案手法で実装したシステムで発見可能な 1 日あたりの隠語の数には実質的な上限がある．この課題を解消する手法として，隠語発見の際に収集したページがアンダーグラウンド系 BBS であるか調べることで，検索 API の利用回数を節約することが考えられる．

これらの研究課題に対しては，今後の研究を通じて取り組んでいく予定である．

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] 2ch 掲示板. 薬、違法板. http://anago.2ch.net/ihou/.
[2] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, Vol. 2, pp. 27:1–27:27, 2011. http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[3] Google. Custom search. https://developers.google.com/custom-search/.
[4] Jetrun テクノロジ株式会社. 隠語・誘導語データベース. http://www.kkyg.jp/.
[5] James Tin-Yau Kwok. Automated text categorization using support vector machine. 1998.
[6] Pinkbbs. お水・出会い系掲示板. http://kilauea.bbspink.com/pub/.
[7] Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In Proceedings 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.
[8] J. J. Rocchio. Relevance feedback in information retrieval. Information Storage and Retrieval: Scientific Report ISR-9, 1965.
[9] 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.
[10] Yahoo!JAPAN. キーフレーズ抽出 api. http://developer.yahoo.co.jp/webapi/jlp/keyphrase/v1/extract.html.
[11] 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.
[12] 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.
[13] 小学館. 日本大百科全書 (ニッポニカ).
[14] 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. Master's thesis, 東京大学大学院 情報理工学系研究科, 2012.
[15] 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.
[16] 柴田武. 現代社会とことば. ことばの講座, No. 5. 東京創元社, 1956.
[17] 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.
[18] 米川明彦. 集団語の研究 (上巻). 東京堂出版, 2009.
[19] 渡辺友左. 階層と言語. 岩波講座日本語, No. 2. 岩波書店, 1977.