# Hidden Web サイトからの新規トピック文書の抽出

毛利隆軌[†,☆] 北川博之[†]

現在インターネット上には様々な Web 情報源が存在している．情報源の中には，問合せインタフェースを介して種々のデータベースコンテンツを提供する情報源が存在する．Hidden Web[2] サイトはそのような情報源の代表的な例である．一方，インターネットが情報流通の基盤となった今日では，Web コンテンツの分析に対するニーズが増大しており，コンテンツの分析による各種の知識発見が要求されている．Hidden Web サイトの情報源が内包するコンテンツも，社会における関心事や情報ニーズを分析する際の手がかりとなる貴重な資源である．特に，新規性の高いトピックの検出やトレンドの分析等の知識発見応用においては，そのコンテンツの時間的変化傾向を知ることが重要となる．しかし，Hidden Web サイトにおいては，利用者がコンテンツ管理者から特別な手助けなしに問合せインタフェースのみを用いてその変化傾向を知ることは一般的に困難である．本論文では，テキストデータベースを内包する Hidden Web サイトが提供する通常のキーワードに基づく問合せインタフェースのみを利用して，テキストデータベース中から新規性の高い文書を重点的に抽出するための手法を提案する．また，実テキストデータを用いた実験を行い，本手法の有効性を評価する．

# Extracting New Topic Documents from Hidden Web Sites

Takanori Mouri[†,☆] and Hiroyuki Kitagawa[†]

There are many information sources which provide their database contents through query interfaces. Hidden web sites are typical examples. Usually, their database contents dynamically change, new documents on emerging topics being appended. In applications like topic detection and trend analysis, it is important to discover newly emerging contents in the databases. However, it is very difficult for ordinary users to detect them only through the query interfaces without support by the database contents administrators. We propose a method to automatically look for such contents in a text database associated with a hidden web site. The proposed method generates biased query probes using a classifier and they are issued to the keyword-based query interface of the given hidden web site. The query probes focus on extracting documents on newly emerging topics. We evaluate effectiveness of the proposed method with intensive experiments.

## 1. はじめに

現在インターネット上には様々な Web 情報源が存在している．情報源の中には，問合せインタフェースを介して種々のデータベースコンテンツを提供する情報源が存在する．Hidden Web[2] サイトはそのような情報源の代表的な例である．一方，インターネットが情報流通の基盤となった今日では，Web コンテンツの分析に対するニーズが増大している．Hidden Web サイト等の情報源が内包するコンテンツも，社会における関心事や情報ニーズを分析する際の手がかりとなる貴重な資源である．特に，新規性の高いトピックの検出やトレンドの分析等の知識発見応用においては，そのコンテンツの時間的変化傾向を知ることが重要となる．

しかし，Hidden Web サイトにおいては，一般の利用者がそのコンテンツアクセスに利用可能な手段は，通常，キーワードに基づく問合せインタフェース等の単純なものに限られており，利用者自身が問合せ条件を工夫して新規性の高いコンテンツを抽出することは非常に困難である．データベースコンテンツ全体をダウンロードできるような状況の場合には，以前のスナップショットと現在のスナップショットを直接比較分析することで変化傾向を知ることが可能である．しかし，このような手段は必ずしもすべての情報源に適用できない．また，それが可能な場合でも，ダウンロードした大量のコンテンツを比較分析するための効率的

† 筑波大学システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba
☆ 現在，株式会社リコーソフトウェア研究開発本部
Presently with RICOH Software R&D Group

な手段が必要となる．

本論文では，テキストデータベースを内包する Hidden Web サイトを対象として，それが提供するキーワードベースの問合せインタフェースのみを利用して，新規性の高いコンテンツを有する文書を重点的に抽出するための手法を提案する．提案方法のポイントは，新規性の高い文書を抽出するのに向いた問合せ用のキーワードを特定する方法である．

本研究に関する予備的検討結果を論文 12)，13) に報告した．論文 12) では，本論文の 3.3 節に示す語選択法の一部について，One-Class SVM[3] に基づく分類器を用いた手法を検討した．また，論文 13) では，論文 12) と同様の語選択法について，One-Class SVM に基づく分類器と階層的クラスタリングに基づく分類器の比較評価を行った．本論文では，新たな語選択法として情報利得[6] を用いた手法を提案し，これまでに検討した語選択法との比較検討を行う．また，論文 12)，13) では，10 トピック，500 件程度の小規模なデータを用いた実験を行っていたのに対し，本論文では 50 トピック，6,000 件程度のより大規模なデータを用い，かつより詳細な検討を行っている．なお，本論文では，論文 13) の評価結果に基づき，階層的クラスタリングを用いた分類手法のみを考察対象としている．

以下では，2 章において関連研究について述べる，3 章では本研究における提案手法を述べる．4 章ではニュースデータを用いた実験について述べ，最後にまとめと今後の課題について述べる．

## 2. 関 連 研 究

### 2.1 Hidden Web サイトに関する研究

本研究が対象とする Hidden Web サイト等のコンテンツの概要を，キーワードに基づく問合せインタフェースのみを用いて抽出するための研究が最近いくつか行われている[1),5)]．これらの方法では，情報源に対して問合せプローブ（query probe）と呼ぶ問合せを多数発行し，サンプル文書を獲得する．これらのサンプル文書から情報源が内包するデータベースのコンテンツを推定する．また，サンプル文書に出現した語やその出現頻度をまとめたものをコンテンツサマリと呼び，当該データベースコンテンツの一種のプロファイルとして用いる．これらの研究は，情報源のコンテンツのある時点でのスナップショットのプロファイルを問合せプローブを用いて獲得することを目的としている．本研究では Hidden Web サイト内の新規性の高いコンテンツの抽出を目的としており，これらの既存研究とは目的が異なる．これらの研究で提案されているようなプロービングを 2 回行い，それぞれで得られるサンプル文書やコンテンツサマリを比較することでコンテンツの変化傾向を分析する方法も考えうる．しかし，本提案の diff プロービングに比べて従来のプロービングには多くの問合せプローブの発行が必要なことや，コンテンツの部分的な変化を多数のサンプル文書や全体的なコンテンツサマリの中から見い出すのは容易でないといった問題点がある．

### 2.2 トピック検出に関する研究

ニュースストリームから新規性の高いトピックを検出する研究として，これまでいくつかの研究が行われている[8),10),11)]．これらの研究では，対象となるニュースデータに対してクラスタリングを行い，新規トピックを有するニュースの発見を行っている．しかし，これらの研究では到着するデータコンテンツをすべて直接的に分析対象とすることが可能な状況を想定している．

テキストデータベースにおけるトレンドの発見を行う試みとして，論文 4) がある．この研究では，テキストデータベース内の文書を直接すべてスキャンすることが可能であり，また文書のタイムスタンプが利用できることを前提としている．

本研究は，Hidden Web サイト等，問合せインタフェースを介してのみコンテンツの抽出が可能な情報源を対象としており，この点で従来のトピック検出等に関する研究が想定している環境とは大きく異なる．

## 3. 提 案 手 法

文書群をコンテンツとし，キーワードに基づく問合せインタフェースを持つテキストデータベース $db$ を内包する Hidden Web サイトが存在するものとする．問合せ結果は何らかの基準でランク付けされて返されるものとする．ある時刻における $db$ のスナップショットを $db(t_1)$ とする．次にこの $db$ に $db(t_1)$ には含まれていなかった新規トピックを含む文書が追加され，時刻 $t_2(> t_1)$ における $db$ のスナップショットが $db(t_2)$ になったとする．本論文では，Hidden Web サイトが受付可能なキーワードを問合せとして発行することにより，$db(t_2) - db(t_1)$ の文書内の新規トピック文書を重点的に抽出するための手法を提案する．

提案手法は，次の 3 つのステップからなる．

**Step 1：初期プロービング**

初期プロービングとは，$db(t_1)$ のコンテンツの情報を取得するための操作である．時刻 $t_1$ において実行される．**初期プローブ**と呼ぶ問合せを情報源に発行し，

$n_1$ 件のサンプル文書（**初期サンプル文書**）を取得する．

**Step 2：分類器の生成**

生成する分類器は与えられた文書が新規トピック文書であるか，そうでないかを判別する分類器である．Step 1 で取得した $n_1$ 件の初期サンプル文書群を基に分類器を生成する．この分類器は，与えられた文書が初期サンプル文書群とどの程度類似しているかを判断し，新規性の高い文書かどうかを判断可能なものとする．

**Step 3：diff プロービング**

diff プロービングとは，新規トピック文書を抽出する操作である．時刻 $t_2$ において実行される．**diff プローブ**と呼ぶ問合せを情報源に発行する．得られた文書を Step 2 で生成した分類器にかけて，新規性の高い文書であると判定された文書のみを**抽出文書**とする．抽出文書数が与えられた件数の $n_2$ になるまで，または新規性の高い文書を取得する割合が小さくなるまでこのステップを実行する．

以下に，各ステップについてより詳細に説明する．

### 3.1 初期プロービング

初期プロービングの手法は，論文 1)，5) で用いられているプロービング手法と同様である．辞書データが利用可能であるものとし，次の 3 つの手順で行う．

1-1 語 $w$ を選択し（詳細は下記参照），データベースに $w$ のみをキーワードとする問合せを発行する．

1-2 問合せ結果から上位最大 $k_1$ 件の文書を取得する．

1-3 取得した文書数が $n_1$ に達した場合終了する．それ以外の場合は手順 1-1 に戻る．

手順 1-1 での語 $w$ の選択の方法は，最初は辞書からランダムに 1 語を取り出す．2 回目以降は，辞書からランダムに取り出す方法（RS-Ord）と，取得した文書内の語からランダムに取り出す方法（RS-Lrd）があげられており，一般的に後者の方が有効であることが示されている[5]．本研究では RS-Lrd を用いる．

### 3.2 分類器の生成

生成する分類器は与えられた文書が新規トピック文書であるか，そうでないかを判別可能な分類器である．生成方法として，本研究では階層的クラスタリング手法[7] を用いる．分類器の生成方法のアルゴリズムは基本的に以下の 3 つの手順で行う．

2-1 初期プロービングにおいて取得した初期サンプル文書中の各文書について，それぞれを 1 文書からなるクラスタと見なす．また，すべてのクラスタの組の類似度を下記に示す余弦尺度を用いて計算する．

2-2 最も類似度が高いクラスタの組を併合し，併合によってできたクラスタと他のクラスタの類似度を再計算する．

2-3 すべてのクラスタ間の類似度が閾値 $\theta$ より小さくなるまで手順 2-2 を繰り返す．

これらの処理を行う際には，初期プロービングにおいて取得した初期サンプル文書群に不要語除去や語幹抽出の処理を行い，$TF \cdot IDF$ の重み付けに基づいてベクトル化を行う．ある文書 $d$ における語 $t$ の出現頻度を $f(d,t)$ とし，その文書 $d$ における総単語数を示した $\sum_{s \in d} f(d,s)$ で正規化を行い $tf(d,t)$ を求める．語 $t$ が出現する文書数を $df(t)$ とおく．このとき，ある文書 $d$ における語 $t$ の重み $w(d,t)$ は

$$w(d,t) = tf(d,t) \cdot idf(t)$$
$$tf(d,t) = \frac{f(d,t)}{\sum_{s \in d} f(s,t)}$$
$$idf(t) = \log \frac{n_1}{df(t)}$$

と与えられる．生成したベクトル間の余弦尺度を類似度としてクラスタを生成していく．

クラスタの併合時には，クラスタ $c_i$ と $c_j$ を併合したクラスタ $c_{ij}$ とクラスタ $c_k$ $(k \neq i, j)$ との類似度 $\theta_{ij,k}$ は次により計算する．

$$\theta_{ij,k} = \frac{1}{2}\theta_{i,k} + \frac{1}{2}\theta_{j,k} - \frac{1}{2}|\theta_{i,k} - \theta_{j,k}|$$

ただし，$\theta_{i,k}$，$\theta_{j,k}$ はそれぞれクラスタ $c_i$ と $c_k$，クラスタ $c_j$ と $c_k$ の間の類似度である．すなわち，2 つのクラスタの最も類似度が小さい文書間の類似度を 2 つのクラスタの類似度とする．

新規トピック文書であるかどうかの判別は次のとおりである．与えられた文書のベクトル化を行い，生成されたすべてのクラスタとの類似度を計算する．すべてのクラスタとの類似度が閾値 $\theta$ よりも小さい場合は新規トピック文書と判別する．いずれかのクラスタとの類似度が閾値 $\theta$ より大きい場合は新規トピック文書ではないと判別する．

**単一文書からなるクラスタの扱い**

閾値によっては，初期サンプル文書をクラスタリングした結果の中に，1 つの文書からなるクラスタが存在することがある．これらの文書はいわゆる外れ値である場合も多く，これらの文書と類似していることを理由に，与えられた文書を新規トピック文書でないと判別すると，判別誤りを生じる場合が多く発生し，新規トピック文書が正しく抽出できなくなる．そこで，上記の判別においては，単一文書からなるクラスタは除外して考える．

**新規に現れる語の $IDF$ 値の計算方法**

類似度を計算するとき，diff プローブによって取得する文書中には初期サンプル文書群には出現しない新たな語 $t_{new}$ が含まれる場合がある．このような語 $t_{new}$ の $IDF$ の値は $df(t)=1$，$n_1 \simeq n_1+1$ と考え

$$idf(t_{new}) = \log(n_1)$$

として重み付けを行う．

### 3.3 diff プロービング

diff プロービングは以下の 4 つの手順で行う．

3-1 語 $w$ を選択し（下記参照のこと），データベースに $w$ のみをキーワードとする問合せ（diff プローブ）を発行する．

3-2 問合せ結果から上位最大 $k_2$ 件の文書**（候補文書）**を取得する．

3-3 取得した最大 $k_2$ 件の候補文書を Step2 で作成した分類器を用いて新規トピック文書であるかどうか判別を行う．新規トピック文書であると判別された場合，その文書を抽出文書に加える．抽出文書数がユーザの欲しいと想定している件数 $n_2$ を満たした場合終了する．

3-4 抽出文書数が $n_2$ に達しない場合，取得した最大 $k_2$ 件の候補文書中の抽出文書の割合を調べる．その割合が $\varepsilon$ 以上の場合は，問合せ結果の次の最大 $k_2$ 件の文書を候補文書として取得し手順 3-3 に戻る．しかし，$\varepsilon$ に満たない場合，あるいは新たに取得可能な文書がない場合は手順 3-1 に戻る．

手順 3-1 における語 $w$ の選択は，最初は初期プロービングと同様に，辞書からランダムに 1 語選ぶものとする．2 回目以降の選択方法として，次の 3 つの方法を比較検討する．

**方法 1：抽出文書中の語のランダムな選択**

抽出文書に含まれる語からランダムに取得する．このとき，頻度は考慮せず異なりに基づいてランダムに選択する．

**方法 2：新出語のランダムな選択**

抽出文書に含まれる語からランダムに選択するが，初期サンプル文書に含まれていた語は除く．すなわち，diff プロービング時に初めて出現した語から選択する．方法 1 と同様に異なりに基づいてランダムに選択する．

**方法 3：情報利得最大語の選択**

抽出文書の集合を $P$ とし，その件数を $p$ とする．また，新規トピック文書でないと判断された候補文書と初期サンプル文書の集合を $Q$ とし，その件数を $q$ とする．文書集合 $P$ 内の文書に出現する語 $w$ について情報利得を調べる．文書集合 $P$ 内の語 $w$ を含む文書数を $p_i$，含まない文書数を $p_j$ とする．また，文書集合 $Q$ 内の語 $w$ を含む文書数を $q_i$，含まない文書数を $q_j$ とする．このとき，語 $w$ の情報利得 $Gain(w)$ は以下の式で求められる．

$$Gain(w) = I(p,q) - E(w)$$

$$E(w) = \frac{p_i+q_i}{p+q} I(p_i,q_i) + \frac{p_j+q_j}{p+q} I(p_j,q_j)$$

$$I(p,q) = -\frac{p}{p+q}\log_2\frac{p}{p+q} - \frac{q}{p+q}\log_2\frac{q}{p+q}$$

このとき，情報利得が最も大きい語 $w$ を次の問合せ語として選択する．すなわち，新規トピック文書でないと判別された文書や初期サンプル文書にあまり含まれず，抽出文書に多く含まれるような語を選択する．

## 4. ニュースデータを用いた実験

### 4.1 実 験 環 境

実際の Hidden Web サイトを用いて本手法を評価することも考えられるが，定量的で客観的な評価が難しいという問題がある．そこで，ある程度内容が分かった文書データを用いて人工的に Hidden Web サイトを構築して，評価を行うこととした．

実験対象の文書データとしては，1998 年の Topic Detection and Tracking（TDT）Phase 2[9] で使われたデータを用いる．これは CNN Headline News や New York Times 等 6 種類の配信源における 1998 年 1 月から 6 月までのニュース記事を集録したコーパスである．集録されたニュース記事の一部にはあらかじめ人手によるトピック付けおよびトピックとの適合度合い（完全に適合するか部分的に適合するかの 2 種類）の情報が付加されている（1 つのニュース記事が属するトピックはたかだか 1 つである）．ここでは，10 件以上の完全に適合するニュース記事を持つトピック 55 個を選択し，それらのトピックラベルが付いた記事を実験に用いる（付録の**表 4**）．

実験では 1 つのニュース記事を 1 文書として扱う．表 4 のすべてのトピックを用いて $db(t_1)$ と $db(t_2)$ となるデータベースを構築した．また，構築した Hidden Web サイトにおけるランキングの処理は，$TF \cdot IDF$ 法を用いた余弦尺度によるものとした．データベースは，新規トピック文書として用いるトピックに属する文書を除いた，$TP_1$ から $TP_{46}$ までの 46 トピックに属する文書を用いてデータベース $db(t_1)$ を構築した．この 46 トピックに属する文書を従来トピック文書と呼ぶ．この $db(t_1)$ に新規トピック文書として新たに文書を追加して $db(t_2)$ を構築した．

実験では階層的クラスタリングの閾値を変化させたとき，抽出した文書中にどのくらい新規トピック文書

が含まれていたかを調べるため，抽出文書中の新規トピック文書の割合を算出した．また，その抽出文書を取得するまでに必要とした diff プローブ数が何回であったかを調べた．

### 4.2 実験内容

#### 4.2.1 実験 1

46 トピックに属する 5,600 件の文書を用いて，データベース $db(t_1)$ を構築した．$db(t_1)$ に新規トピック文書として 1 つのトピックに属する文書を追加して実験を行った．新規トピックとして追加するトピックを変化させて実験を行った．

新規トピックとして追加するトピック文書は，$db(t_1)$ の全文書数と比べて，(1) 約 0.076 の割合となる $TP_{47}$ に属する 427 件，(2) 約 0.096 の割合となる $TP_{48}$ に属する 540 件，(3) 約 0.05 の割合となる $TP_{49}$ に属する 280 件，(4) 約 0.036 となる $TP_{50}$ に属する 203 件，(5) 約 0.025 の割合となる $TP_{51}$ に属する 126 件の 5 種類のトピック文書である．実験は各トピックごとに行った．各トピックに属する文書を新規トピック文書として追加して $db(t_2)$ を構築した．

#### 4.2.2 実験 2

実験 1 と同様に 46 トピックに属する 5,600 件の文書を用いて，データベース $db(t_1)$ を構築した．本実験では $db(t_1)$ に新規トピック文書として 4 つのトピックに属する文書を追加して $db(t_2)$ を構築した．新規トピックとして追加するトピックの組合せを変化させて実験を行った．新規トピック文書として追加する文書は，$db(t_1)$ の全文書数の 0.08 の割合にあたる 448 件（各トピック 112 件）の文書である．追加する新規トピックは，(1) $TP_{47}$，$TP_{50}$，$TP_{51}$，$TP_{53}$，(2) $TP_{48}$，$TP_{49}$，$TP_{54}$，$TP_{55}$，(3) $TP_{47}$，$TP_{48}$，$TP_{49}$，$TP_{50}$，(4) $TP_{51}$，$TP_{53}$，$TP_{54}$，$TP_{55}$，(5) $TP_{47}$，$TP_{50}$，$TP_{54}$，$TP_{55}$，の 5 つの組合せである．

本実験においては，抽出文書中の新規トピック文書の割合と抽出文書を取得するまでに必要な diff プロービング数だけでなく，抽出文書中の新規トピック文書が属するトピックの種類についても調べた．

#### 4.2.3 実験 3

実際のデータベースでは更新されるとき，新規トピックに関する文書だけではなく，更新前のデータベースに含まれていたトピックに属する文書，すなわち従来トピック文書も加わることが予想される．本実験では $db(t_2)$ を構築する際に，$db(t_1)$ に追加する文書として新規トピック文書だけでなく，従来トピック文書も加えた場合について実験を行った．本実験では，実験 1，2 で構築していた $db(t_1)$ の $TP_1$ から $TP_{46}$ の 46 トピックに属する全文書から各トピック 8 割ずつ，4,480 件の文書を用いて $db(t_1)$ を構築した．$db(t_1)$ に追加する新規トピック文書としては，実験 1-(1) で用いたトピック $TP_{47}$ に属する文書を用い，従来トピック文書としては $db(t_1)$ を構築する際に用いなかった $TP_1$ から $TP_{46}$ の残りの 2 割の文書を用いた．すなわち追加する文書は，新規トピックとして 427 件の文書と，従来トピック文書として $db(t_1)$ に存在する 46 トピックと同じトピックに属する 1,120 件の文書である．合計 1,547 件の文書を $db(t_1)$ に追加して $db(t_2)$ を構築した．評価としては，$TP_{47}$ に属する文書 427 件のみを新規トピック文書として扱った．

#### 4.2.4 実験 4

実験 1，2，3 では初期サンプル文書数を 300 件と固定していた．初期サンプル文書数を変化させると，分類器の精度が変化することが予想される．よって初期サンプル文書数を 100 件，300 件，500 件，700 件と変化させて実験を行った．

そのときの新規トピック文書の割合と diff プロービング数について調べた．またクラスタリングによってできるクラスタ数について調べ，各クラスタの主要トピックを求め，全体としてクラスタがいくつトピックをカバーしているかについて調べた．また生成されるクラスタの純度についても調べた．クラスタの純度は，クラスタ内の文書中で最も多いトピックを主要トピックとし，その主要トピックのクラスタ内の割合を算出し，その平均とした．

実験環境としては実験 1-(1) のトピック $TP_{47}$ を用いた場合と同様である．

**パラメータの設定**

実験に用いたパラメータを**表 1** に示す．$k_1$，$n_1$ の値は論文 1)，5) より，Hidden Web サイトが持つテキストデータベースのコンテンツサマリを表現するのに十分であると述べられている値に基づきこの値を用いた．$k_2$ の値は，Yahoo [14]，Google [15]，ODP [16] 等のサイトの問合せ結果が 1 ページごとに 10 件または 20 件表示することから，本実験では 20 を用いた．

### 4.3 実験結果

実験は 3 種類の diff プロービングの方法を用いて行った．実験結果は 50 回測定した値の平均値を示している．各回でプロービング開始時に辞書から選択する語は異なる．

閾値 $\theta$ は 0.04，0.08，0.12，0.16 で行った．

#### 4.3.1 実験 1

**新規トピック文書割合**

抽出文書 40 件の中の新規トピック文書の割合につ

**表 1** パラメータ
Table 1 Parameters.

| パラメータ | 変数 | 値 |
|---|---|---|
| 初期プロービング時に取得する文書数 | $k_1$ | 4 |
| diff プロービング時に取得する文書数 | $k_2$ | 20 |
| 初期サンプル文書数 | $n_1$ | 300 |
| 抽出文書数 | $n_2$ | 40 |
| 新たな問合せを発行するための閾値 | $\varepsilon$ | 0.5 |
| 階層的クラスタリングの閾値 | $\theta$ | 0.04, 0.08, 0.12, 0.16 |

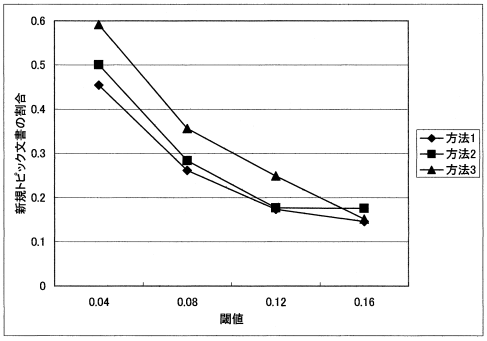



**図 1** 実験 1-(1) の結果：新規トピック文書割合
Fig. 1 Experiment 1-(1): The rate of new topic documents in the extracted documents.

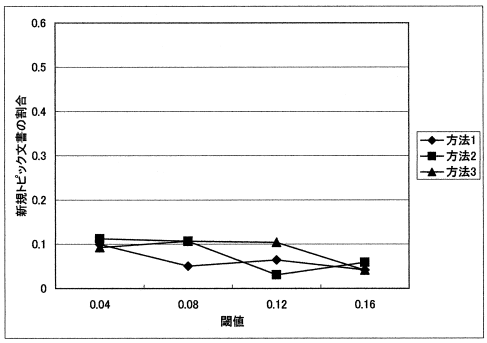



**図 2** 実験 1-(4) の結果：新規トピック文書割合
Fig. 2 Experiment 1-(4): The rate of new topic documents in the extracted documents.

いて調べた．最も新規トピック文書の割合が高い実験 1-(1) と最も低い実験 1-(4) の実験結果を**図 1**，**図 2** に示す．他の実験結果は付録の**表 5** に示す．

実験 1-(1) の結果である図 1 より，閾値 $\theta = 0.04$ のとき，方法 3 では，新規トピック文書の割合は約 0.59 であり，方法 2 では約 0.50 である．また，方法 1 では約 0.45 の割合で新規トピック文書を抽出することができた．このデータベース中に存在する新規トピック文書の割合は 0.08 以下であることを考慮すると，完全にランダムなキーワードを用いてサンプリングした場合は約 0.08 以下の割合でしか取得できないと考えられる．これに対して，方法 3 ではこの約 7 倍，方法 2 では約 6 倍，方法 1 では約 5.5 倍の精度で新規トピック文書が抽出できている．

図 2 で分かるとおり，実験 1-(4) に関しては他のトピックを加えた実験と異なり，いずれの閾値においても抽出文書中の新規トピック文書の割合が低くなってしまった．割合が低くなってしまったと考えられる理由については後で述べる．

実験結果より，ほとんどの場合において閾値 $\theta = 0.04$ の場合が最も新規トピック文書の割合が高く，閾値 $\theta = 0.16$ の場合が最も新規トピック文書の割合が低い．これは，閾値が小さくなると，新規トピック文書であると判別する基準が厳しくなることによる．また閾値が大きくなると新規トピック文書の割合が下がるのは，逆に判別する基準が緩くなることによる．

実験結果より，3 種類の diff プロービングのうちで高い割合で新規トピック文書を取得できたのは，方法 2 と方法 3 である．この 2 つの方法は取得した文書からより新規性の高い文書を取得するようなプローブの語を選択しているので高い割合で新規トピック文書が抽出することができたと考えられる．

**分類時の誤り率**

表 5 より実験 1-(4) が他の実験と比べて，新規トピック文書の割合が低いことが分かる．ここで最も新規トピック文書の割合が高い実験 1-(1) と実験 1-(4) の分類器の誤り率を調べた．diff プローブで抽出文書を 40 件抽出するまでに取得した候補文書全体を分類する際，従来トピック文書を新規トピック文書として誤った割合と，新規トピック文書を従来トピック文書として誤った割合について調べた．

実験 1-(1) における従来トピック文書を新規トピック文書として誤った割合を**図 3** に示し，新規トピック文書を従来トピック文書として誤った割合を**図 4** に示す．実験 1-(4) における従来トピック文書を新規トピック文書として誤った割合を**図 5** に示し，新規トピック文書を従来トピック文書として誤った割合を**図 6** に示す．

実験 1-(1) と実験 1-(4) において，分類器を生成するために用いた文書はともに $db(t_1)$ に含まれる文書

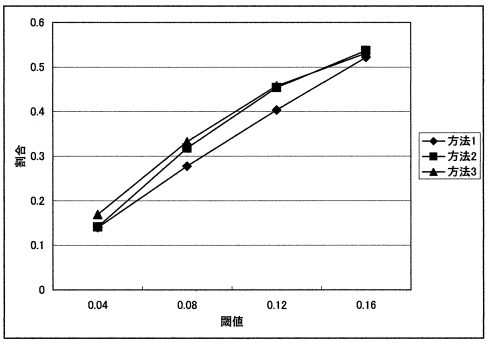



**図 3** 実験 1-(1) の結果：従来トピック文書を新規トピック文書と誤った割合
Fig. 3 Experiment 1-(1): The error rate of misjudging an old document as a new one.

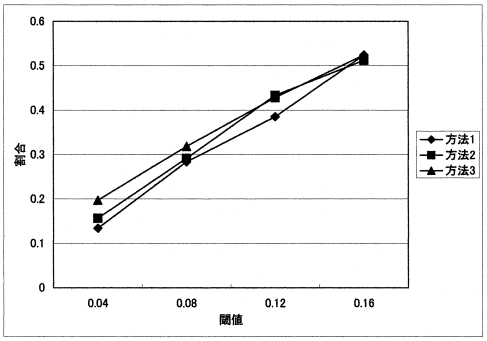



**図 5** 実験 1-(4) の結果：従来トピック文書を新規トピック文書と誤った割合
Fig. 5 Experiment 1-(4): The error rate of misjudging an old document as a new one.

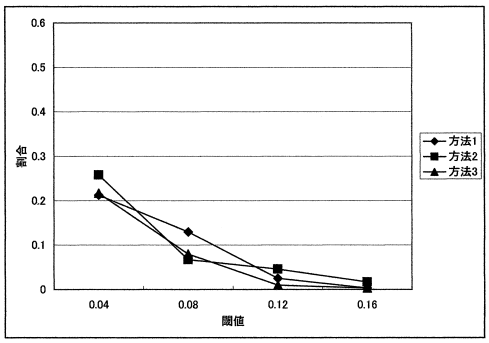



**図 4** 実験 1-(1) の結果：新規トピック文書を従来トピック文書と誤った割合
Fig. 4 Experiment 1-(1): The error rate of misjudging a new document as an old one.

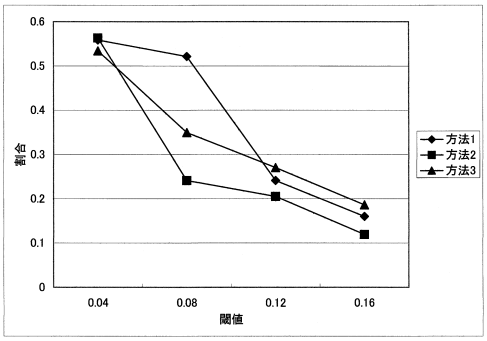



**図 6** 実験 1-(4) の結果：新規トピック文書を従来トピック文書と誤った割合
Fig. 6 Experiment 1-(4): The error rate of misjudging a new document as an old one.

である．図 3 と図 5 から，従来トピック文書を新規トピック文書として誤った割合に大きな違いは見られない．しかし図 4 と図 6 から，新規トピック文書を従来トピック文書として誤った割合は実験 1-(4) が大きいことが分かる．これは実験 1-(4) で用いた新規トピック文書と従来トピック文書との類似度が高く，分類器が誤りやすい文書であったと予想される．よって今後はこのような従来トピック文書との類似度が高いトピック文書においても抽出できるような仕組みを考える必要がある．

**diff プローブ数**

抽出文書を 40 件抽出するまでに発行した diff プローブ数を調べた．

実験 1-(1) の結果を**図 7**，実験 1-(4) の結果を**図 8** に示す．他の実験結果は付録の**表 6** に示す．いずれの diff プロービングの方法においても，閾値が小さい場合がプローブ数が多く，閾値が大きい場合がプローブ数が少ない．同様の傾向は，実験を行ったいずれの場合においても見られた．閾値が小さくなると，新規トピック文書であると判別する際の基準が厳しくなる．したがって，既存のクラスタとの類似度が低い文書を得るために多くの文書を取得することになり，diff プローブ数が多くなってしまうと考えられる．

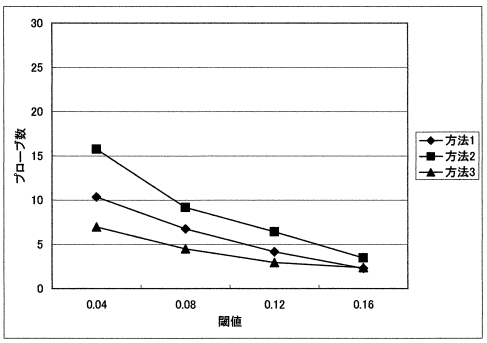



**図 7** 実験 1-(1) の結果：diff プローブ数
Fig. 7 Experiment 1-(1): The number of diff probes.

また，いずれのトピックを用いた実験においても，

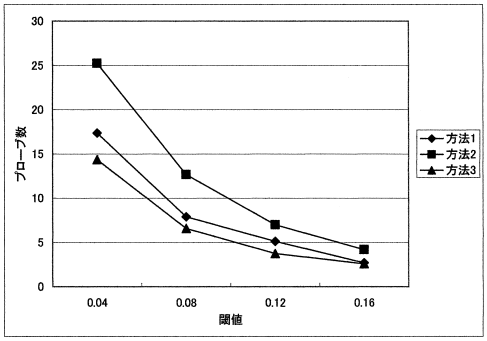



**図 8** 実験 1-(4) の結果：diff プローブ数
Fig. 8 Experiment 1-(4): The number of diff probes.

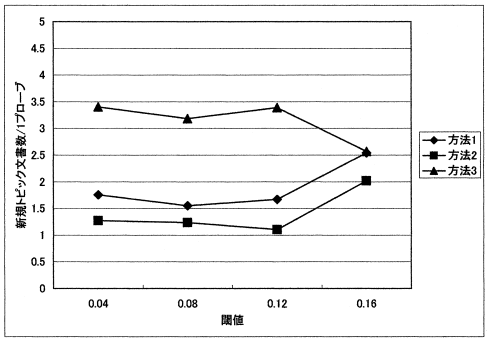



**図 9** 実験 1-(1) の結果：1 diff プローブあたりの新規トピック文書抽出数
Fig. 9 Experiment 1-(1): The number of new topic documents per a diff probe.

方法 3 が最も少ないプローブ数で必要数の抽出文書を得ることができている．

新規トピック文書の割合を示した結果と diff プローブ数を示した結果から考えられるのは，新規トピック文書の割合と，プローブ数にはトレードオフの関係があるということである．そこで，次に実験 1-(1)，実験 1-(4) において，1 回の diff プローブあたり平均何件の新規トピック文書を抽出したかをそれぞれ調べた．

実験 1-(1) の結果を**図 9** に，実験 1-(4) の結果を**図 10** に示す．図より方法 3 が他の方法より 1 回の diff プローブでより多くの新規トピック文書を抽出できていることが分かる．他の実験においても同様の結果を算出できる．

これより，方法 3 が他の方法より優れており，最も効率良く新規トピック文書を抽出していることが分かる．

### 4.3.2 実 験 2

**新規トピック文書割合**

結果は付録の**表 7** に示す．実験結果から実験 2 でも実験 1 と同じ傾向が見られた．閾値 $\theta = 0.04$ の場合

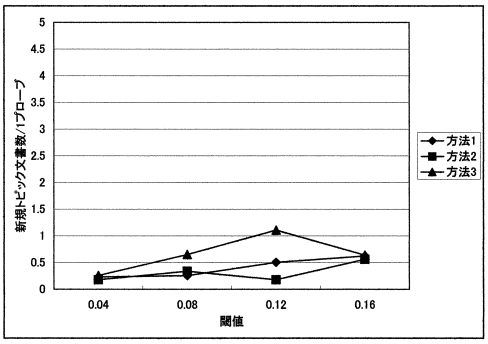



**図 10** 実験 1-(4) の結果：1 diff プローブあたりの新規トピック文書抽出数
Fig. 10 Experiment 1-(4): The number of new topic documents per a diff probe.

は新規トピック文書の割合が高く，閾値 $\theta = 0.16$ の場合は新規トピック文書の割合が低くなった．また実験 2 では実験 1 と異なり，新規トピックとして 4 つのトピックが新たに加わっている．このことから複数のトピックが加わった場合においても本手法は有効であるといえる．方法 2 の新規トピック文書の割合が比較的高いのは，以下に述べる抽出したトピック数に関する性質が影響している可能性がある．

**diff プローブ数**

抽出文書を 40 件抽出するまでに発行した diff プローブ数を調べた．

結果は付録の**表 8** に示す．diff プローブ数に関しても，実験 1 と同様の結果が得られ，閾値が小さくなると diff プローブ数が増え，閾値が大きくなると diff プローブ数が少なくなった．

よって実験 1 と同様に，新規トピック文書の割合とプローブ数にはトレードオフの関係があることが分かる．3 つの方法の中では方法 3 が最も効率良く新規トピック文書を抽出できるといえる．

**抽出したトピック数**

本実験では複数のトピック文書を追加している．そこで，本実験では抽出したトピック数について調べた．すなわち，1 回の実験において，抽出した 40 件の抽出文書中に何トピックが含まれているかを調べた．この場合，多くの新規トピック文書が抽出された場合について調べる方が本手法の傾向が分かると思われる．よって最も多くの新規トピック文書が抽出できた実験 2-(4) の $\theta = 0.04$ の場合について調べた．その結果を**図 11** に示す．横軸が 1 回の実験で抽出した新規トピック数を示し，縦軸が 50 回の実験のうち，その新規トピック数を抽出した回数を示している．

いずれの方法においても，全 4 種類の新規トピック

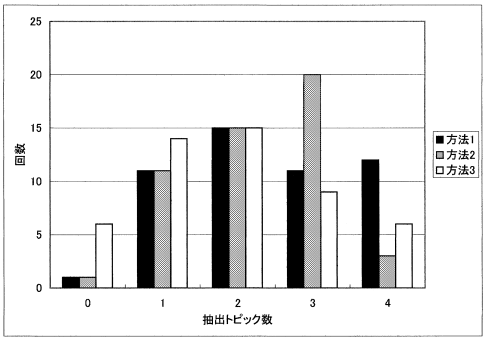



**図 11** 実験 2 の結果：抽出したトピック数
Fig. 11 Experiment 2: The number of extracted topics.

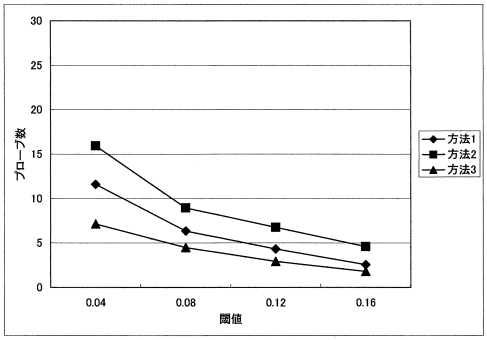



**図 13** 実験 3 の結果：diff プローブ数
Fig. 13 Experiment 3: The number of diff probes.

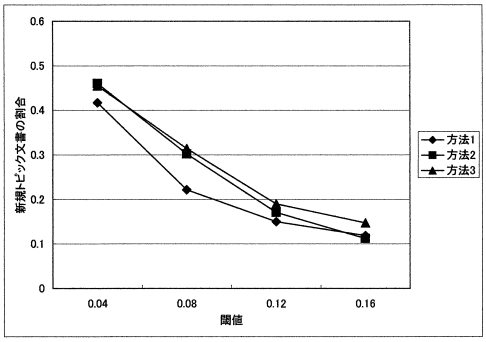



**図 12** 実験 3 の結果：新規トピック文書割合
Fig. 12 Experiment 3: The rate of new topic documents in the extracted documents.

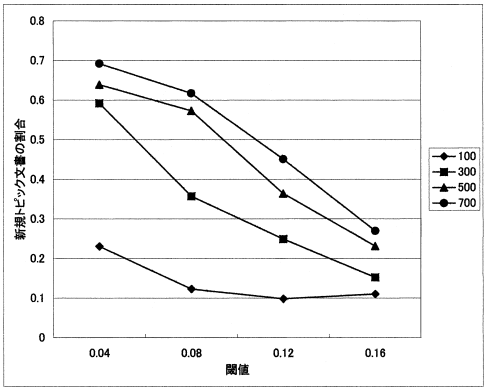



**図 14** 実験 4 の結果：新規トピック文書割合
Fig. 14 Experiment 4: The rate of new topic documents in the extracted documents.

文書を抽出していることは少ない．方法 1 が 4 種類のトピックを抽出していることが多く，3 つの方法の中では比較的多くのトピックを 1 回の実験で取得している．次に 3，4 種類のトピックを多く取得しているのは方法 2 である．方法 3 は 0，1，2 種類のトピックを取得していることが多い．方法 3 ではある 1 つのトピックに属する文書を抽出すると，次の diff プローブ用の語もそのトピックに属する文書を取得するような語になる傾向があり，それによってこのような結果が生じていると考えられる．一方，方法 1，方法 2 ではランダムに diff プローブ用の語が選択されるので，同じ種類のトピックだけでなく，複数のトピックが抽出される傾向があると考えられる．

#### 4.3.3 実　験　3
#### 実験 1 との比較

実験の結果を**図 12**，**図 13** に示す．新規トピック文書として同じトピック文書を扱い，なおかつ従来のトピックを追加していない場合である実験 1-(1) と比較する．図 1 と図 7 が実験 1-(1) の結果である．新規トピック文書の割合において，閾値 $\theta = 0.04$ のときの方法 3 の新規トピック文書の割合が少し下がっている．しかし他の場合においては大きな変化がほとんど見られず，同じ傾向が見られた．diff プローブ数に関してはほとんど同じ結果であった．よって本手法は新規トピック文書だけでなく，従来トピック文書を新たに加えた場合においても，新規トピック文書を効率良く抽出することができることがいえる．

#### 4.3.4 実　験　4
#### 新規トピック文書の割合と diff プローブ数

実験は 50 回行った．初期サンプル文書数を変化させたときの新規トピック文書の割合と diff プローブ数の結果を，付録の**表 9** と**表 10** に示す．全体的に新規トピック文書が多く抽出できている方法 3 について，新規トピック文書の割合と diff プローブ数を**図 14** と**図 15** に示す．図より初期サンプル文書数を増加させると新規トピック文書の割合が高くなることが分かる．しかし，diff プローブ数も増加していることも分かる．最も高い割合で新規トピック文書が取得できている閾値 0.04 において，初期サンプル文書数が 300 件の 0.592 と比べると，500 件，700 件では 0.639，0.692 と割合が増加していることが分かる．diff プローブ数

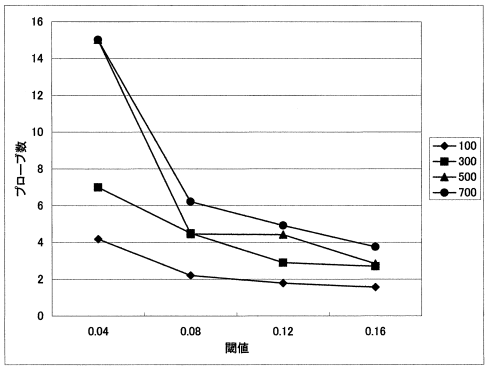



**図 15** 実験 4 の結果：diff プローブ数
Fig. 15 Experiment 4: The number of diff probes.

は初期サンプル文書数が 300 件の場合の 7 回と比べて 500 件，700 件の場合は 15 回と，ともに 8 回増加していることが分かる．これは，初期サンプル文書数の増加により分類器の精度が向上し，従来トピック文書を新規トピック文書と誤って判定するケースが減ることによると考えられる．

1 回の diff プローブあたり抽出した新規トピック文書数の平均を調べると，300 件では 3.4 件，500 件では 1.7 件，700 件では 1.8 件となっている．本実験において閾値 0.04 の場合，初期サンプル文書数が 300 件のとき最も効率良く抽出できているといえる．閾値が 0.08 の場合では初期サンプル文書数が 500 件の場合が最も効率が良い．また方法 1，方法 2 を用いた実験結果でも，比較的 300 件が効率が良い場合が多い．これより，初期サンプル文書を増加させると新規トピック文書の割合は増加するが，1 回の diff プローブの効率が落ちることが分かる．

初期サンプル文書数を 100 件から 300 件に変化させた場合と比べて，500 件から 700 件と変化させた場合は，新規トピック文書数の割合の変化は小さい．また，初期サンプル文書数が 700 件の場合，46 トピック中 40 トピック抽出でき，ほとんどのトピックが抽出できている．よって初期サンプル文書数を 700 件より大きくした場合，増加させた初期サンプル文書数の割合に対して新規トピック文書の割合の増加はさらに小さくなると思われる．

**トピック数とクラスタ数**

初期サンプル文書を変化させたとき，生成されるクラスタがカバーしているトピック数の平均を**表 2** に示す．またクラスタ数の平均とクラスタの純度の平均を**表 3** に示す．

表 2 より，初期サンプル文書数が多くなるとカバーするトピック数も増加することが分かる．初期状態で

**表 2** 実験 4 の結果：クラスタが対応しているトピック数
Table 2 Experiment 4: The number of topics.

| 初期サンプル文書数 | 閾値 | | | |
|---|---|---|---|---|
| | 0.04 | 0.08 | 0.12 | 0.16 |
| 100 | 14.9 | 14.2 | 13.2 | 12.8 |
| 300 | 29.9 | 29.9 | 29.1 | 28.5 |
| 500 | 34.6 | 34.7 | 34.5 | 34.0 |
| 700 | 40.0 | 40.4 | 40.3 | 40.0 |

**表 3** 実験 4 の結果：クラスタ数とクラスタの純度
Table 3 Experiment 4: The number of the clusters and the purity of the clusters.

| 初期サンプル文書数 | 調査内容 | 閾値 | | | |
|---|---|---|---|---|---|
| | | 0.04 | 0.08 | 0.12 | 0.16 |
| 100 | クラスタ数 | 26.2 | 27.7 | 26.9 | 25.5 |
| | 純度 | 0.87 | 0.92 | 0.96 | 0.97 |
| 300 | クラスタ数 | 63.0 | 74.8 | 81.0 | 126.2 |
| | 純度 | 0.89 | 0.92 | 0.94 | 0.95 |
| 500 | クラスタ数 | 85.2 | 105.9 | 118.6 | 126.2 |
| | 純度 | 0.91 | 0.93 | 0.95 | 0.96 |
| 700 | クラスタ数 | 105.8 | 136.5 | 156.9 | 170.1 |
| | 純度 | 0.91 | 0.93 | 0.94 | 0.96 |

のデータベースは全部で 46 トピックなので，初期サンプル文書が 700 件の場合は，ほとんどのトピックをカバーしていることが分かる．カバーしているトピック数の増加によって新規トピック文書の割合が高くなったと考えられる．しかし閾値を変化させた場合については，カバーするトピック数はあまり変化が見られなかった．

表 3 よりクラスタの純度はいずれの場合においても高い割合を示している．したがって生成されたクラスタはそれぞれ 1 つのトピックに対応しているといえる．またクラスタ数に関して，初期サンプル文書数が少ない場合はクラスタ数が少なくなる．これは初期サンプル文書数が少ない場合，除外する 1 件の文書からなるクラスタが多く生成されるので，実際のクラスタ数は少なくなる．また初期サンプル文書数が増加した場合，ある 1 つのトピックに関するクラスタが複数個生成されるのでクラスタ数が多くなる．

#### 4.3.5 実験のまとめ

実験より 3 種類の diff プロービングの中で方法 3 が高い割合で新規トピック文書を抽出でき，diff プローブ数が少なかった．1 diff プローブあたりの新規トピック文書抽出数においても方法 3 が最も高く，最も効率良く新規トピック文書を抽出できることが示された．方法 3 は，情報利得最大語を用いているので，新規トピックの特徴となる語が diff プローブ用の語として選択される．そのため，新規トピック文書を重点的に取得することができ，新規トピック文書の割合が高くな

ると考えられる．しかし，複数のトピックが追加された場合，平均的にすべてのトピックを 1 回の操作で抽出することは少ない．この問題を解決するには，ある 1 種類のトピックの文書を必要数抽出できたら，そのトピックを新規トピックではなく，従来トピックとして扱う等して分類器や diff プローブの仕組みを変更する方法が考えられる．

実験 3 より，データベース更新時に新規トピック文書だけでなく従来トピック文書が追加された場合においても，本手法が有効であることが分かった．

実験 4 より，初期サンプル文書数を変化させるとクラスタ数が増加し，カバーするトピック数も増加する．また新規トピック文書の割合が増加することが分かった．しかし初期サンプル文書数を増加させると，diff プローブ数も増加するため，効率良く文書を抽出するには適当ではないことが分かった．

また初期サンプル文書の数にかかわらず，生成されたクラスタは純度が高く，ある 1 つのトピックを表しているといえる．

## 5. まとめと今後の課題

本研究では，動的に新規トピック文書が追加更新されるテキストデータベースを内包する Hidden Web サイトから，新規性の高い文書を抽出するための手法を提案した．3 種類の diff プロービングを実験により比較し，方法 3 の情報利得を用いた diff プロービングが最も効率良く新規トピック文書を抽出することができることを示した．

今後の課題として，4 章で述べた，抽出しにくいトピック文書の抽出方法の開発，複数の新規トピックを平均的に抽出する方法の開発や，日本語文書を用いた実験があげられる．また時間情報をもとに文書を追加していった場合の実験もあげられる．さらに本手法で得た抽出文書から新規トピックそのものを抽出する方法についても検討が必要である．

また本手法は，Hidden Web サイトにおける新規トピック文書の抽出のみではなく，複数のテキストデータベースコンテンツの差分情報の抽出等にも応用することができると考えられる．そのような視点からの検討も今後必要である．

**謝辞** 本研究の一部は，科学研究費補助金基盤研究（B）(#15300027) および特定領域研究（2）(#16016205) の助成による．

## 参 考 文 献

1) Callan, J. and Connell, M.: Query-Based Sampling of Text Databases, *ACM TOIS*, Vol.19, No.2 (2001).
2) Florescu, D., Levy, A.Y. and Mendelzon, A.O.: Database techniques for the World-Wide Web: A survey, *SIGMOD Record*, Vol.27, No.3, pp.59–74 (1998).
3) Maevitz, L.M. and Yousef, M.: One-Class SVMs for Document Classification, Journal of Machine Learning Research, Vol.2, No.2, pp.139–154 (2002).
4) Brian, L., Agrawal, R. and Srikant, R.: Discovering Trends in Text Databases, *Proc. 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining*, Newport Beach, California (August 1997).
5) Ipeirotis, P.G. and Gravano, L.: Distributed Search over the Hidden Web: Hierarchical Database Sampling and Selection, *Proc. 28th VLDB Conf.* (2002).
6) Quinlan J.R: *C4.5: Programs For Machine Learning*, Morgan Kaufmann Publishers Inc. (1993).
7) Salton, G.: *Automatic Information Organization and Retrieval*, McGraw-Hill Book Company (1968).
8) Topic Detection Task. http://www.nist.gov/speech/tests/tdt/tasks/detect.htm
9) 1998 Topic Detection and Tracking Project (TDT-2). http://www.nist.gov/speech/tests/tdt/tdt98/
10) Walls, F., Jin, H., Sista, S. and Schwartz, R.: Topic detection in broadcast news, *Proc. DARPA Broadcast News Workshop*, pp.193–198, Morgan Kaufmann Publishers, Inc., San Francisco, CA (1999).
11) Yang, Y., Zhang, J., Carbonell, J. and Jin, C.: Topic-conditioned Novelty Detection, *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.688–693 (2002).
12) 毛利隆軌，北川博之：プロービングによるテキストデータベースからの新規トピック文書抽出，日本データベース学会 Letters, Vol.2, No.1, pp.107–110 (2003).
13) 毛利隆軌，北川博之：プロービングとクラスタリングによる新規トピック文書抽出，日本データベース学会 Letters, Vol.2, No.2, pp.76–80 (2003).
14) http://www.yahoo.com/
15) http://www.google.com/
16) http://www.dmoz.org/

## 付録 実験に用いたトピックと実験結果

**表 4** TDT データのトピックラベルと文書数
Table 4 TDT data set.

| トピック ID | トピックラベル | 文書数 |
|---|---|---|
| $TP_1$ | Current Conflict with Iraq | 1322 |
| $TP_2$ | Asian Economic Crisis | 1032 |
| $TP_3$ | Monica Lewinsky Case | 969 |
| $TP_4$ | Anti-Suharto Violence | 324 |
| $TP_5$ | Pope visits Cuba | 151 |
| $TP_6$ | Violence in Algeria | 125 |
| $TP_7$ | Unabomber | 119 |
| $TP_8$ | Anti-Chinese Violence in Indonesia | 36 |
| $TP_9$ | Afghan Earthquake | 23 |
| $TP_{10}$ | German Train derails | 51 |
| $TP_{11}$ | Puerto Rico phone strike | 12 |
| $TP_{12}$ | Clinton-Jiang Debate | 68 |
| $TP_{13}$ | McVeigh's Navy Dismissal & Fight | 19 |
| $TP_{14}$ | Upcoming Philippine Elections | 41 |
| $TP_{15}$ | Fossett's Balloon Ride | 15 |
| $TP_{16}$ | Casey Martin Sues PGA | 56 |
| $TP_{17}$ | Karla Faye Tucker | 48 |
| $TP_{18}$ | State of the Union Address | 42 |
| $TP_{19}$ | Babbitt Casino Case | 20 |
| $TP_{20}$ | Bombing AL Clinic | 99 |
| $TP_{21}$ | Cable Car Crash | 110 |
| $TP_{22}$ | China Airlines Crash | 36 |
| $TP_{23}$ | Tornado in Florida | 53 |
| $TP_{24}$ | Diane Zamora | 30 |
| $TP_{25}$ | Shevardnadze Assassination Attempt | 38 |
| $TP_{26}$ | Oprah Lawsuit | 70 |
| $TP_{27}$ | Mary Kay LeTourneau | 12 |
| $TP_{28}$ | John Glenn | 37 |
| $TP_{29}$ | Superbowl '98 | 84 |
| $TP_{30}$ | David Satcher confirmed | 16 |
| $TP_{31}$ | Quality of Life, NYC | 33 |
| $TP_{32}$ | Grossberg baby murder | 26 |
| $TP_{33}$ | Asteroid Coming?? | 31 |
| $TP_{34}$ | Dr. Spock Dies | 15 |
| $TP_{35}$ | Viagra Approval | 93 |
| $TP_{36}$ | JJ the Whale | 11 |
| $TP_{37}$ | James Earl Ray's Retrial? | 49 |
| $TP_{38}$ | World Figure Skating Champs | 20 |
| $TP_{39}$ | Bird Watchers Hostage | 16 |
| $TP_{40}$ | Race Relations Meetings | 12 |
| $TP_{41}$ | Rats in Space! | 60 |
| $TP_{42}$ | Tony Awards | 14 |
| $TP_{43}$ | Nigerian Protest Violence | 50 |
| $TP_{44}$ | Denmark Strike | 15 |
| $TP_{45}$ | World AIDS Conference | 18 |
| $TP_{46}$ | NBA finals | 79 |
| $TP_{47}$ | India, A Nuclear Power? | 427 |
| $TP_{48}$ | 1998 Winter Olympics | 540 |
| $TP_{49}$ | National Tobacco Settlement | 280 |
| $TP_{50}$ | Israeli-Palestinian Talks (London) | 203 |
| $TP_{51}$ | Sgt. Gene McKinney | 126 |
| $TP_{52}$ | Pope visits Cuba | 151 |
| $TP_{53}$ | GM Strike | 138 |
| $TP_{54}$ | Jonesboro shooting | 125 |
| $TP_{55}$ | India Parliamentary Elections | 121 |

**表 5** 実験 1 の結果：新規トピック文書の割合
Table 5 Experiment 1: The rate of new topic documents in the extracted documents.

| 実験 | 方法 | 閾値 | | | |
|---|---|---|---|---|---|
| (トピック ID) | | 0.04 | 0.08 | 0.12 | 0.16 |
| 1-(1) | 方法 1 | 0.455 | 0.262 | 0.174 | 0.146 |
| ($TP_{47}$) | 方法 2 | 0.501 | 0.284 | 0.177 | 0.176 |
| | 方法 3 | 0.592 | 0.357 | 0.249 | 0.152 |
| 1-(2) | 方法 1 | 0.313 | 0.369 | 0.155 | 0.150 |
| ($TP_{48}$) | 方法 2 | 0.521 | 0.160 | 0.242 | 0.160 |
| | 方法 3 | 0.406 | 0.373 | 0.200 | 0.141 |
| 1-(3) | 方法 1 | 0.319 | 0.204 | 0.207 | 0.116 |
| ($TP_{49}$) | 方法 2 | 0.329 | 0.219 | 0.114 | 0.091 |
| | 方法 3 | 0.336 | 0.272 | 0.184 | 0.119 |
| 1-(4) | 方法 1 | 0.100 | 0.051 | 0.065 | 0.042 |
| ($TP_{50}$) | 方法 2 | 0.113 | 0.107 | 0.031 | 0.059 |
| | 方法 3 | 0.093 | 0.107 | 0.104 | 0.042 |
| 1-(5) | 方法 1 | 0.285 | 0.110 | 0.057 | 0.056 |
| ($TP_{51}$) | 方法 2 | 0.283 | 0.152 | 0.049 | 0.037 |
| | 方法 3 | 0.284 | 0.213 | 0.117 | 0.066 |

**表 6** 実験 1 の結果：diff プローブ数
Table 6 Experiment 1: The number of diff probes.

| 実験 | 方法 | 閾値 | | | |
|---|---|---|---|---|---|
| (トピック ID) | | 0.04 | 0.08 | 0.12 | 0.16 |
| 1-(1) | 方法 1 | 10.7 | 6.8 | 4.2 | 2.3 |
| ($TP_{47}$) | 方法 2 | 15.8 | 9.2 | 6.4 | 3.5 |
| | 方法 3 | 7.0 | 4.5 | 2.9 | 2.7 |
| 1-(2) | 方法 1 | 14.6 | 5.9 | 3.9 | 2.6 |
| ($TP_{48}$) | 方法 2 | 15.0 | 9.2 | 5.4 | 3.3 |
| | 方法 3 | 11.2 | 4.8 | 3.1 | 2.4 |
| 1-(3) | 方法 1 | 14.2 | 6.8 | 4.4 | 2.7 |
| ($TP_{49}$) | 方法 2 | 19.8 | 9.7 | 7.3 | 3.9 |
| | 方法 3 | 11.3 | 4.5 | 3.7 | 2.4 |
| 1-(4) | 方法 1 | 17.4 | 7.9 | 5.1 | 2.7 |
| ($TP_{50}$) | 方法 2 | 25.2 | 12.7 | 7 | 4.2 |
| | 方法 3 | 14.4 | 6.6 | 3.8 | 2.6 |
| 1-(5) | 方法 1 | 13.8 | 7.4 | 4.5 | 2.5 |
| ($TP_{51}$) | 方法 2 | 22.5 | 12.0 | 7.4 | 4.3 |
| | 方法 3 | 12.9 | 4.7 | 3.4 | 2.4 |

**表 7** 実験 2 の結果：新規トピック文書の割合
Table 7 Experiment 2: The rate of new topic documents in the extracted documents.

| 実験（トピック ID） | 方法 | 閾値 | | | |
|---|---|---|---|---|---|
| | | 0.04 | 0.08 | 0.12 | 0.16 |
| 2-(1) | 方法 1 | 0.334 | 0.230 | 0.166 | 0.181 |
| （$TP_{47}$, $TP_{50}$, $TP_{51}$, $TP_{53}$） | 方法 2 | 0.415 | 0.297 | 0.179 | 0.180 |
| | 方法 3 | 0.394 | 0.316 | 0.198 | 0.102 |
| 2-(2) | 方法 1 | 0.318 | 0.231 | 0.167 | 0.164 |
| （$TP_{48}$, $TP_{49}$, $TP_{54}$, $TP_{55}$） | 方法 2 | 0.436 | 0.357 | 0.290 | 0.212 |
| | 方法 3 | 0.417 | 0.328 | 0.161 | 0.161 |
| 2-(3) | 方法 1 | 0.162 | 0.189 | 0.101 | 0.114 |
| （$TP_{47}$, $TP_{48}$, $TP_{49}$, $TP_{50}$） | 方法 2 | 0.429 | 0.333 | 0.229 | 0.129 |
| | 方法 3 | 0.331 | 0.291 | 0.121 | 0.100 |
| 2-(4) | 方法 1 | 0.427 | 0.267 | 0.216 | 0.167 |
| （$TP_{51}$, $TP_{53}$, $TP_{54}$, $TP_{55}$） | 方法 2 | 0.463 | 0.334 | 0.318 | 0.218 |
| | 方法 3 | 0.436 | 0.333 | 0.321 | 0.197 |
| 2-(5) | 方法 1 | 0.307 | 0.252 | 0.140 | 0.113 |
| （$TP_{47}$, $TP_{50}$, $TP_{54}$, $TP_{55}$） | 方法 2 | 0.440 | 0.335 | 0.230 | 0.139 |
| | 方法 3 | 0.308 | 0.275 | 0.176 | 0.167 |

**表 8** 実験 2 の結果：diff プローブ数
Table 8 Experiment 2: The number of diff probes.

| 実験（トピック ID） | 方法 | 閾値 | | | |
|---|---|---|---|---|---|
| | | 0.04 | 0.08 | 0.12 | 0.16 |
| 2-(1) | 方法 1 | 13.2 | 6.9 | 4.2 | 2.2 |
| （$TP_{47}$, $TP_{50}$, $TP_{51}$, $TP_{53}$） | 方法 2 | 17.1 | 9.9 | 5.9 | 3.7 |
| | 方法 3 | 8.3 | 4.5 | 3.1 | 2.1 |
| 2-(2) | 方法 1 | 14.7 | 7.1 | 4.2 | 2.4 |
| （$TP_{48}$, $TP_{49}$, $TP_{54}$, $TP_{55}$） | 方法 2 | 16.5 | 9.2 | 6.1 | 3.7 |
| | 方法 3 | 9.4 | 4.5 | 3.3 | 2.1 |
| 2-(3) | 方法 1 | 15.7 | 8.2 | 3.9 | 2.7 |
| （$TP_{47}$, $TP_{48}$, $TP_{49}$, $TP_{50}$） | 方法 2 | 18.7 | 8.5 | 5.2 | 3.4 |
| | 方法 3 | 11.1 | 5.2 | 3.4 | 2.2 |
| 2-(4) | 方法 1 | 12.4 | 6.7 | 3.8 | 2.3 |
| （$TP_{51}$, $TP_{53}$, $TP_{54}$, $TP_{55}$） | 方法 2 | 15.9 | 8.6 | 4.8 | 4.1 |
| | 方法 3 | 7.9 | 4.8 | 2.9 | 2.0 |
| 2-(5) | 方法 1 | 15.3 | 7.0 | 4.4 | 2.6 |
| （$TP_{47}$, $TP_{50}$, $TP_{54}$, $TP_{55}$） | 方法 2 | 17.6 | 10.2 | 6.2 | 2.9 |
| | 方法 3 | 11.2 | 5.3 | 3.4 | 2.3 |

**表 9** 実験 4 の結果：新規トピック文書の割合
Table 9 Experiment 4: The rate of new topic documents in extracted documents.

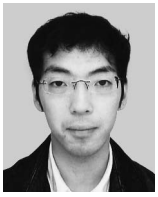
| 初期サンプル文書数 | 方法 | 閾値 | | | |
|---|---|---|---|---|---|
| | | 0.04 | 0.08 | 0.12 | 0.16 |
| 100 | 方法 1 | 0.150 | 0.120 | 0.96 | 0.07 |
| | 方法 2 | 0.193 | 0.091 | 0.096 | 0.104 |
| | 方法 3 | 0.231 | 0.123 | 0.098 | 0.110 |
| 300 | 方法 1 | 0.455 | 0.262 | 0.174 | 0.146 |
| | 方法 2 | 0.501 | 0.284 | 0.177 | 0.176 |
| | 方法 3 | 0.592 | 0.357 | 0.249 | 0.152 |
| 500 | 方法 1 | 0.489 | 0.307 | 0.172 | 0.140 |
| | 方法 2 | 0.599 | 0.433 | 0.336 | 0.172 |
| | 方法 3 | 0.639 | 0.573 | 0.364 | 0.231 |
| 700 | 方法 1 | 0.609 | 0.394 | 0.297 | 0.191 |
| | 方法 2 | 0.741 | 0.578 | 0.416 | 0.308 |
| | 方法 3 | 0.692 | 0.617 | 0.451 | 0.270 |

**表 10** 実験 4 の結果：diff プローブ数
Table 10 Experiment 4: The number of diff probes.

| 初期サンプル文書数 | 方法 | 閾値 | | | |
|---|---|---|---|---|---|
| | | 0.04 | 0.08 | 0.12 | 0.16 |
| 100 | 方法 1 | 5.8 | 2.7 | 1.6 | 1.4 |
| | 方法 2 | 6.5 | 3.5 | 1.7 | 1.6 |
| | 方法 3 | 4.2 | 2.2 | 1.8 | 1.6 |
| 300 | 方法 1 | 10.7 | 6.8 | 4.2 | 2.3 |
| | 方法 2 | 15.8 | 9.2 | 6.4 | 3.5 |
| | 方法 3 | 7.0 | 4.5 | 2.9 | 2.7 |
| 500 | 方法 1 | 15.8 | 9.4 | 6.6 | 4.0 |
| | 方法 2 | 22.9 | 14.3 | 12.3 | 7.2 |
| | 方法 3 | 15.0 | 4.5 | 4.4 | 2.8 |
| 700 | 方法 1 | 19.1 | 11.6 | 7.4 | 5.0 |
| | 方法 2 | 28.5 | 21.7 | 18.8 | 14.7 |
| | 方法 3 | 15.0 | 6.2 | 4.9 | 3.8 |

(平成 16 年 9 月 20 日受付)
(平成 17 年 1 月 19 日採録)

(担当編集委員 波多野 賢治)



**毛利 隆軌**(正会員)

2002 年筑波大学第三学群情報学類卒業．2004 年同大学院システム情報工学研究科修士号取得．現在，株式会社リコーソフトウェア研究開発本部勤務．ACM SIGMOD 日本支部，日本データベース学会各会員．

**北川 博之**(正会員)

1978 年東京大学理学部物理学科卒業．1980 年同大学院理学系研究科修士課程修了．日本電気(株)勤務の後，1988 年筑波大学電子・情報工学系講師．同助教授を経て，現在，筑波大学大学院システム情報工学研究科，計算科学研究センター教授．理学博士(東京大学)．異種情報源統合，XML とデータベース，WWW の高度利用，データマイニング等の研究に従事．著書『データベースシステム』(昭晃堂)，『The Unnormalized Relational Data Model』(共著，Springer-Verlag)等．ACM SIGMOD 日本支部長．日本データベース学会理事．ACM，IEEE-CS，電子情報通信学会，日本ソフトウェア科学会各会員．