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

Extracting New Topic Documents by Probing Text Databases and Clustering

毛利 隆軌♥ 北川 博之♦

Takanori MOURI Hiroyuki KITAGAWA

Hidden Web サイトをはじめとして，内包するデータベースコンテンツを問合せインタフェースを介して外部の利用者に提供する情報源が増加している．多くの情報源では，そのコンテンツは時間と共に動的に追加更新される．データベースコンテンツの変化内容を知ることは，新規トピック検出やトレンド分析等の情報利用において重要である．しかし，上記のような情報源においては，利用者がコンテンツ管理者からの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に困難である．我々は，キーワードに基づく問合せインタフェースを有しそのコンテンツが動的に追加更新されるテキストデータベースから，新たに追加された新規性の高いトピックを有するコンテンツを抽出するための手法を提案してきた．しかし，複数のトピックが混在する状況における詳細な手法の検討が不十分であった．そこで本論文では，階層的クラスタリング手法を用いることで，複数のトピックが混在する状況においてより高い割合で新規性の高いトピックを有するコンテンツを抽出する手法について検討を行う．また，実テキストデータを用いた実験により，本手法の有効性を示す．

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, we want 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 proposed a method to automatically discover such content. The proposed method generates biased query probes using a classifier to be issued to a given text database with a keyword-based query interface. In this paper, we enhance the method using a hierarchical clustering to cope with cases that multiple topics are mixed in the database contents. We evaluate its effectiveness with experiments on real text databases.

♥ 学生会員 筑波大学大学院システム情報工学研究科
tmouri@kde.is.tsukuba.ac.jp
♦ 正会員 筑波大学電子・情報工学系
kitagawa@is.tsukuba.ac.jp

1. はじめに

現在，インターネット上には問合せインタフェースを介して様々なデータベースコンテンツを提供する情報源が存在している．Hidden Webサイト等はそのような情報源の代表的な例である．インターネットが情報流通の基盤となった今日では，これらの情報源が内包するコンテンツは，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，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一般に非常に困難である．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，このような手段が全ての情報源に適用できる訳ではない．また，それが可能であっても，大量のコンテンツをダウンロードし比較分析するための効率的な手段が必要となる．

我々はテキストデータベース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，新規性の高いコンテンツ（文書）を重点的に抽出するための手法を提案してきた[10]．

その手法は，更新する前のコンテンツの情報を取得して，そのコンテンツの情報から分類器を生成し，データベースが更新された後に，その分類器と問合せプローブを用いて新規性の高い文書を抽出するものである．

その手法は，複数のトピックが混在する状況において，クラスタリングのアプローチを取るとしてきたが，その詳細な手法の検討が不十分であった．そこで，本論文では，複数のトピックが混在するコンテンツを持つ場合を対象とした手法として，階層的クラスタリング手法を用いた手法を提案する．また，提案手法の精度を実験を用いて評価する．

以下の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改善した提案手法を述べる．４章ではCNNニュースデータを用いて改善前の手法との比較を行い本手法の有効性を示す．最後にまとめと今後の課題について述べる．

2. 関連研究

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][2]．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．本論文で提案する手法では，３章に述べるように，初期プロービングとdiffプロービングの２段階のプロービングを行う．初期プロービングは，基本的には上記の手法に基づくものであるが，diffプロービングにおいては，新規性が高い文書を抽出する

ための問合せであるdiffプローブを発行する点が特徴である．[1][2]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得る．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出すのは容易でないといった問題点がある．

新規性の高いトピックの検出に関しては，これまでトピック検出等の領域で多くの研究が行われている[5][7][8][9]．これらでは，ニュースストリーム等から新規性の高いトピックを自動的に検出する方法が検討されている．[7][8][9]等の研究では，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全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．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. 提案方式

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もつテキストデータベース db 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．２つの時刻 t_1, t_2 ($t_1 < t_2$)における db のスナップショットを $db(t_1)$, $db(t_2)$とする．本論文では，db が処理可能な問合せを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を有する文書をより多く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．

Step1: 初期プロービング

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することを，n_1 のサンプル文書(**初期サンプル文書**)を取得するまで繰り返す．

Step 2: クラスタの作成

Step 1 で取得した n_1 件の初期サンプル文書に対して階層的クラスタリング手法を用いてクラスタの生成を行う．

Step 3: diff プロービング

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．得られた文書と Step 2 で生成した各クラスタとの類似度を計算してどのクラスタに属さないと判定された文書のみを**抽出文書**とする．抽出文書数が n_2 件となるまで，この操作を繰り返す．

以下に，各ステップのより詳細について説明する．

3.1 初期プロービング

初期プロービングの手法は，[1][2]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) 語 w を選択し(詳細は下記)，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．
(2) 問合せ結果から上位 k 件の文書を取得する．
(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)に戻る．

手順(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)が挙げられており，一般的に後者の方が有効であるが示されている[2]. 本研究では RS-Lrd を用いる．

3.2 クラスタの作成

クラスタの生成手法として，本研究では階層的クラスタリング手法[4]を用いる．アルゴリズムは基本的に以下の 3 つの手順で行う．

(1) 各文書だけから成るクラスタを生成する．すべてのクラスタの組の類似度を余弦尺度を用いて計算する．
(2) もっとも類似度が高いクラスタの組を併合する．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計算する．
(3) すべてのクラスタ間の類似度が閾値 θより小さくなるまで(2)を繰り返す．

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った後，TF・IDF の重み付けに基づいてベクトルを生成する．ある文書 d における語 t の重み w(d,t)は

$$w(d,t) = tf(d,t) \cdot idf(t)$$

$$tf(d,f) = \frac{f(d,t)}{\sum_{s \in d} f(d,s)}$$

$$idf(t) = \log \frac{n_1}{df(t)}$$

と与えられる．生成したベクトルを基に類似度を計算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．

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

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で 2 つのクラスタの類似度を近似する．

3.3 diff プロービング

diff プロービングは以下の 3 つの手順で行う．

(1) 語 w を選択し(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(diff プローブ)を発行する．
(2) 問合せ結果から上位 k 件の文書(**候補文書**)を取得する．
(3) (2)で取得した k 件の候補文書を Step2 で作成した各クラスタとの類似度を調べる．

どのクラスタに対してもその文書との類似度が閾値 θ以上にならない場合，新規文書と判断してその文書を抽出文書に加える．抽出文書数が n_2 に達した場合終了する．それ以外の場合は手順(1)に戻る．

手順(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，[10]で次の 3 つの方法を提案してきた．

方法１．抽出文書に含まれる語からランダムに取得する．

方法２．抽出文書に含まれる語からランダムに選択するが，分類を行った際，抽出文書に含めるべきでないと判断された候補文書に含まれていた語は除く．

方法 3．抽出文書に含まれる単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．

Step2 において，閾値によっては 1 つの文書からなるクラスタが多く存在することとなる．クラスタリングで併合を行う条件が，クラスタ内の文書との類似度のうち最も小さい値が閾値より高いことより，文書が 1 つのクラスタが多く存在すると，併合条件が緩くなり，新規文書であってもいずれかのクラスタと併合する可能性が高くなる．本実験では，前述の処理 Step2 において，初期サンプル文書群より作成したクラスタ群の内，1 つの文書のみからなるクラスタは除いて(3)の処理を行った．

4. CNN ニュースデータを用いた実験

4.1 実験内容

実験対象の文書データとして利用したのは 1998 年の Topic Detection and Tracking (TDT) Phase 2[6]で使われたデータであり，これは CNN Headline News のほか New York Times や AP など 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トピック付けおよび記事とトピックとの適合の具合（完全に適合するか一部のみ適合するかの 2 種類）の情報が付加されている．ここではトピックと完全に適合するニュース記事を選び，そのうち 10 個のトピックを選んで実験を行った(表 1)．実験では 1 つのニュース記事を 1 文書として扱う．これらのデータを基に実験における $db(t_1)$と $db(t_2)$となるデータベースを構築した．また，テキストデータベースの問合せ処理は，TF・IDF 法を用いた余弦尺度によるものとした．

実験

複数のトピックが混在するデータベースに，新たなトピックを 1 つ追加した場合において，閾値を変化させて，候補文書数と抽出文書に含まれる新規文書割合を調べた．また新規文書でないと判断された候補文書数とその候補文書中に含まれていた新規文書数を調べた．さらにクラスタリングを行わずに One Class Support Vector Machine(以下 SVM)[3]を用いる手法[10]によって行ったときの実験結果と比較を行った．

実験には TP_1 から TP_{10} までのデータを使用した．この 10 トピックの文書のうち 1 月から 3 月の記事 475 件の文書を $db(t_1)$とし，4 月から 6 月までの記事 94 件の文書を加えた計 569 件の文書を $db(t_2)$とした(表 1)．本実験では $db(t_1)$には現れず $db(t_2)$のみに現れる TP_{10} に属する文書だけを新規文書として扱うこととする．TP_{10} の文書は 48 件であるので新規文書の割合は全体の 0.084 となる．

パラメータの設定

プロービングを行う際に取得する文書数 k は 4，初期サンプル文書数 n_1 は 300 件とした．これらの k や n_1 の値は文献[2]における実験結果の考察に基づく．抽出文書数 n_2 の値は 30 とした．類似度を調べるための閾値 θを 0.02 から 0.2 まで 0.02 刻みで実験を行った．

4.2 実験結果

実験の結果を図 1 に示す．10 回の実験結果の平均を表している．また本手法による改善を行う前の結果と本手法で得られた結果を比較したものを図 3 に示す．実線は抽出文書 30 件中の新規文書の割合を示し，破線は 30 件を抽出するまでに取得した候補文書数を示している．

図 1 より，閾値 θが小さくなる程新規文書割合が高くなることが分かる．θ= 0.02 の場合を考えると新規文書の割合は 0.7 になる．これはデータベース全体の新規文書の割合の

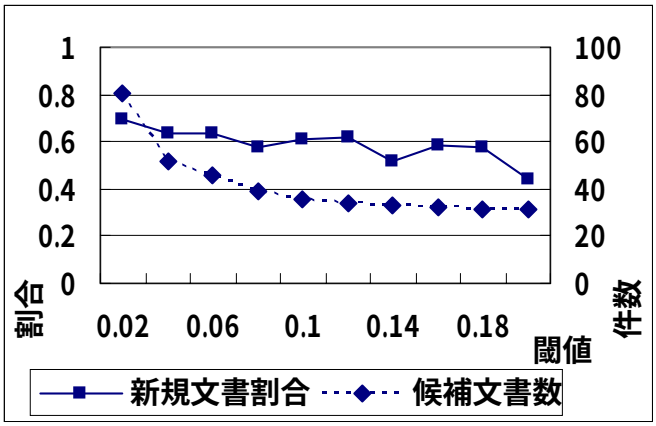

図 1　実験 1 に用いたトピックと文書数

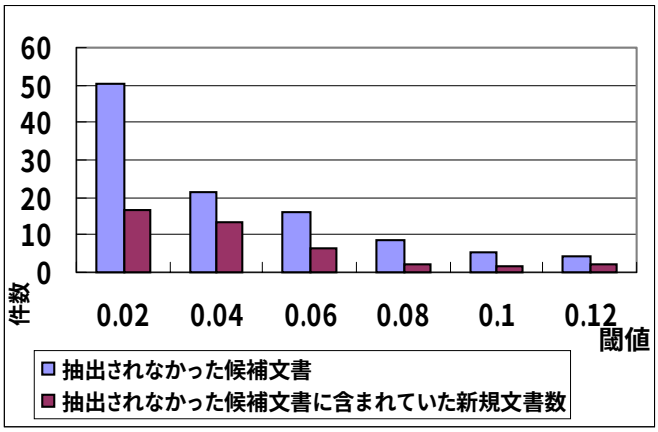

図 2　実験 1 に用いたトピックと文書数

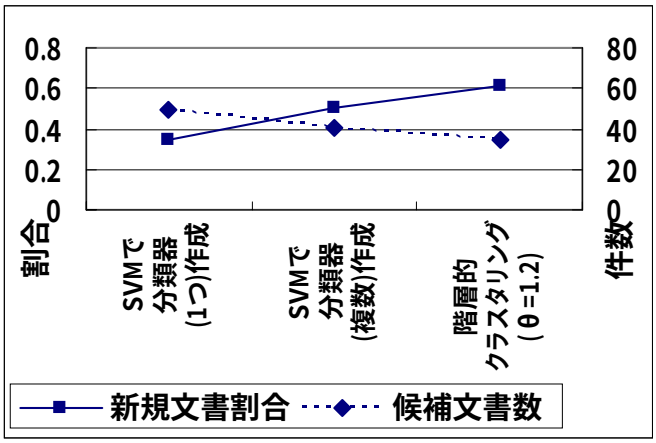

図 3　実験 1 に用いたトピックと文書数

0.084 を考えると，ランダムにサンプリングする場合と比べて，8 倍の精度で新規文書を抽出することができていると言える．グラフから 0.02 から 0.2 までの範囲では，ほとんどの状況で抽出文書に含まれる新規文書割合が 5 割以上であることが分かる．この結果から本手法の有効性が確認できる．

次に候補文書数を考える．図 2 に示したのは，30 件の抽出文書数を得るまでに新規文書ではないと判断された文書数とその文書内の含まれていた新規文書の数を表している．θ= 0.02 の場合では，50.1 件の文書が新規文書ではないと判

断されている．さらにこの新規文書でないと判断された文書群中に存在する新規文書数は 16.7 件となる．これは小さい閾値の場合，本来新規文書である文書を新規文書でないと誤って判断する場合が増えることを表す．次に θ= 0.12 の場合を調べると，候補文書数が 34.3 件となっており 4.3 件の文書が新規でないと判断されている．新規文書でないと判断される文書中に存在する本来新規文書である文書数は 1.9 件となっている．また本実験において，閾値 θ= 0.1 以上からは候補文書数が 30 件から 40 件の間となり，新規文書でないと判断される文書中に存在する本来新規文書である件数は少なくなると予想される．

以上のことから閾値を小さくすると新規文書の割合が増えるが，候補文書数が大きくなる．また候補文書中にある新規文書を新規文書ではないと誤って判断することが多くなる．逆に閾値を大きくすると新規文書の割合は減るが，候補文書数が減り，本来新規文書である文書を新規文書でないと誤って判断することが少なくなると言える．

次に候補文書数が比較的少なくて新規文書割合の高い閾値 θ=0.12 の場合について，[10]で述べた手法と比較を行った(図 3)．横軸における右の項目は初期サンプル文書群を一つのクラスタとして扱い，そのクラスタから SVM を用いて一つの分類器を生成して diff プロービングによって文書を抽出した結果である．真ん中の項目は，初期サンプル文書群をトピックラベルに基づき 9 つのクラスタに分類して各クラスタに対して分類器を生成し，diff プローブによって文書を抽出した結果である．右の項目が本手法による実験結果である．複数の分類器を生成した手法と本手法を比較すると，グラフから抽出文書中の新規文書の割合が 0.5 から 0.62 となり，本手法が最も新規文書の割合が高いことが分かる.さらに候補文書数も 40 件から 34 件と少なくなっていることが分かる．よって本手法を用いることで手法の改善を行うことができたと言える．

Topic ID	トピック名	初期文書数	追加数
TP_1	アジア経済危機	55	35
TP_2	アラバマ病院爆破事件	62	11
TP_3	ローマ法王のキューバ訪問	35	0
TP_4	長野オリンピック	81	0
TP_5	フロリダのトルネード被害	36	0
TP_6	Diane Zamora への有罪判決	23	0
TP_7	Oprah Winfrey に対しての訴訟	59	0
TP_8	Gene McKinney 軍曹の性的不品行に対する公判	91	0
TP_9	スーパーボール	33	0
TP_{10}	バイアグラ	0	48

表１ CNN データのトピックとデータ数

5. まとめと今後の課題

本研究では，複数のトピックが混在するコンテンツを持つテキストデータベースが追加更新された時，階層的クラスタリング手法を用いて新規性の高い文書を抽出するための手法について検討を行った．CNN のニュースデータに対して実験を行い，クラスタリングを行わない方法に比べて提案手法の優位性を示すことができた．

今後の課題として，問合せ語の選択方法，より高い割合で新規文書を抽出する方法の検討，また実在する Hidden Web サイトを対象とした実験が挙げられる，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，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用い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)，特定領域(2)(15017207)による．

[文献]

[1] J. Callan and M. Connell. Query-Based Sampling of Text Databases. *ACM TOIS* 19(2) 2001

[2] Panagiotis G. Ipeirotis and Luis Gravano. Distributed Search over the Hidden Web: Hierarchical Database Sampling and Selection. *Proc. 28th VLDB Conf.,* 2002.

[3] Larry M.Maevitz and Malik Yousef. One-Class SVMs for Document Classification.

[4] G.Salton. Automatic Information Organization and Retrieval, McGraw-Hill Book Company,1968.

[5] Topic Detection Task. http://www.nist.gov/speech/tests/tdt/tasks/detect/htm.

[6] 1998 Topic Detection and Tracking Project (TDT-2) http://www.nist.gov/speech/tests/tdt/tdt98/

[7] F. Walls, H. Jin, S. Sista, and R. Schwartz. Topic detection in broadcast news. In Proceedings of the DARPA Broadcast News Workshop, pages 193-198, San Francisco, CA, 1999. Morgan Kaufmann Publishers, Inc.

[8] Y. Yang, T. Pierce, and J. Carbonell. A study on retrospective and on-line event detection. In Proceedings of the 21th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'98).page 28-36,1998.

[9] Y. Yang, J. Zhang, J. Carbonell and C. Jin. Topic-conditioned Novelty Detection. ACM SIGKDD Internaltional Conference on Knowledge Discovery and Data Mining, pp 688-693, 2002.

[10] 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 年 5 月.

毛利 隆軌 Takanori MOURI

筑波大学大学院システム情報工学研究科在学中．2002 年筑波大学第三学群情報学類卒業．XML，WWW，文書データベースに興味を持つ．情報処理学会学生会員．日本データベース学会学生会員．

北川 博之 Hiroyuki KITAGAWA

筑波大学電子・情報工学系教授．1980 年東京大学大学院理学系研究科修了．理学博士（東京大学）．異種情報源統合，文書データベース，WWW の高度利用等の研究に従事．著者「データベースシステム」（昭晃堂），「Unnormalized Relational Data Model」(共著，Springer-Verlag)等. ACM, IEEE-CS，日本データベース学会，情報処理学会，電子情報通信学会，日本ソフトウェア科学会，各会員．