逆トピックワードを利用した外れ値文書検出

真下飛瑠
茨城大学工学部情報工学科

新納浩幸
茨城大学工学部情報工学科

佐々木稔
茨城大学工学部情報工学科

1　はじめに

文書セットに対してクラスタリング等の解析を行う場合，外れ値にあたる文書を予め検出しておくことは重要である．ここではある事柄に関連した文書セットを対象に，外れ値文書の検出を行う．

外れ値文書の検出を行う場合，データマイニング分野の外れ値検出の手法を利用するのが一般的である．また外れ値検出手法は多数存在する [4]．しかし，どの手法にも共通した問題点として，ある同類の外れ値がある程度の個数出現した場合，それらを外れ値として検出できないということがある．通常，外れ値は他のデータからの距離が離れていたり，外れ値近辺の密度が低かったり，外れ値の生成確率が低いなどの性質があり，外れ値検出手法はそれらを手がかりに外れ値を検出する．ところが，ある同類の外れ値がある程度の個数出現した場合，上記のいずれの性質も満たさない状況が生じるために，検出に失敗する．

本論文では上記の問題点の解決を主な目的とする．そのためにここでは Bekkerman が提案した関連文書検出の手法を応用する [2]．Bekkerman は対象の文書セットとは別に一般の巨大コーパスを用いて，文書セットの話題と関連度の高いトピックワードを選出し，そのトピックワードを利用して文書セットから関連度の高い文書を順に検出する手法を提案した．ここでは，その手法の一般の巨大コーパスと文書セットを逆に設定することで，一般の文書では頻出するが対象の文書セットでは現れづらい逆トピックワードを選出することで，外れ値文書検出を行う．

実験では，一般の巨大コーパスとして新聞記事３年分を利用する．ここからあるキーワードに関連した文書セットを取り出し，それらに人為的に外れ値文書を混在させ，対象の文書セットからそれら外れ値文書を検出できるかどうかの実験を行った．提案手法の他，外れ値検出手法として LOF [3] と One Class SVM [1] を用い，提案手法の有効性を示す．

2　トピックワードによる関連文書検出

Bekkerman が提案した関連文書検出の手法について述べる．Bekkerman はこの手法を One Class Clustering (OCC) と呼んでいるので，ここでも OCC と呼ぶことにする．

OCC ではある目的で集められた文書セットには，あるトッピックが存在し，そしてその文書セット内の単語が文書のトピックを負う役割を持つと考えた．その役割が高い単語を トピックワード と呼んでいる．

OCC ではトピックワードを利用して関連文書を検出するが，OCC を実現するには以下の問題を解決する必要がある．

1. どのような単語をトピックワードとするか
2. どうやってトピックワードを選出するか
3. そのトピックワードを使い，どのように関連文書を検出するか

2.1　指標 ρ の導入

問題点の 1，2 については，単語 w のトピックへの関連度を示す指標として以下で定義される ρ を導入する．

$$\rho(w) = \frac{p(w)}{q(w)}$$

ここで $p(w)$ は対象の文書セット D 中での w の出現確率，$q(w)$ は一般の巨大コーパス中での w の出現確率を表す．

ρ の値は一般には使われにくく，しかし対象の文書セット中では頻出する単語において高くなる．よって対象文書セット D が十分に大きい場合，$\rho(w)$ の値の高い単語 w をトピックワードとして選出する．

2.2 Max-KL アルゴリズム

問題点の 3 については基本的に Max-KL アルゴリズムを用いる．

Max-KL アルゴリズムは関連文書を見つけるためのアルゴリズムである．文書 p と q と間の関連度を以下の式により測る．

$$KL(p \parallel q) = \sum_{d \in D w \in G} p(d,w) log \frac{p(w)}{q(w)}$$

ここで D は文書セット全体，G は文書セットに含まれる全単語である．この値を利用して関連文書を検出する．

しかし Max-KL アルゴリズムでは均一性をもった文書にしか使えない．そこで OCC では以下の手順で関連文書を検出する．

1. 対象文書セットの全単語に対して $\rho(w)$ を計算する
2. $\rho(w)$ の値が高い単語上位 m 個をトピックワードとする
3. 各文書をトピックワードからなる単語集合として表現しなおす
4. 各文書の KL ダイバージェンスを計算する
5. KL ダイバージェンスの高いものを関連文書とする

KL ダイバージェンスの計算には，次式を使う．R はトピックワードの集合である．

$$KL_d^r(p \parallel q) = \sum_{w \in R} p'(d,w) log \frac{p(w)}{q(w)}$$

$$p'(d,w) = \frac{p(d,w)}{\sum_{w \in R} p(d,w)}$$

3 逆トピックワードの設定

指標 ρ は単語のトピックへの関連度を表すが，ρ を計算する際の「一般の巨大コーパス」と「与えられた文書セット」を逆にすることを考える．つまり以下の値を考える．

$$\rho(w) = \frac{q(w)}{p(w)}$$

この値が高い単語は，一般的な文書には頻出するが，文書セット中で使われづらい単語である．上記の ρ の値が高い単語をここでは 逆トピックワード と呼ぶことにする．

本論文では OCC で使われた ρ を上記の ρ で置き換える，つまりトピックワードを逆トピックワードに替えて，OCC を実行することを提案する．本手法によって文書セットのトピックとは関連度の低い文書が検出できる．

例として，今回の実験で用いた「災害・死者」文書セットにおけるトピックワードと逆トピックワードの各上位 10 個を表 1 に示す．

表 1: 「災害・死者」文書セットの (逆) トピックワード

トピックワード	逆トピックワード
倒	表示
検視	戦
流言	改革
ＦＥＭＡ	大会
神戸海洋気象台	派
烈震	監督
高潮	連続
ノースリッジ	女子
圧死	系
がけ崩れ	捜査

4 外れ値文書検出

本論文では，提案手法の有効性を示すために，一般の外れ値検出手法として LOF と One Class SVM を試し，本手法と比較する．本章ではこれら手法について概説する．

4.1 LOF

外れ値検出の手法は距離ベースの手法，密度ベースの手法，クラスタリングベースの手法等に分類できるが，LOF は密度ベースの代表的な手法である [3]．概略，データの近傍の密度を利用することで，そのデータの外れ値の度合いを測り，その値によって外れ値を検出する．

LOF におけるデータ $x \in D$ における外れ値の度合いを $LOF(x)$ と表記する．ここで D はデータ全体の集合である．$LOF(x)$ を定義するために，いくつかの式を定義しておく．まず $kdist(x)$ は x に対する k 距離と呼ばれる値で，以下の条件を満たすデータ $o \in D$ との距離 $d(x,o)$ として定義される．

1. 少なくとも k 個のデータ $o' \in D \setminus \{x\}$ に対して $d(x,o') \leq d(x,o)$が成立する．
2. 高々$k-1$ 個のデータ $o' \in D \setminus \{x\}$ に対して

$d(x, o') < d(x, o)$が成立する．

直感的には，上記のデータ o はデータ x からの k 番目に近いデータとなる．データ x から同じ距離を持つデータが複数存在する場合を考慮して，上記のようなテクニカルな定義になっている．

次に $kdist(x)$ を利用して，$N_k(x)$，$rd_k(x, y)$ 及び $lrd_k(x)$ を以下のように定義する．

$$N_k(x) = \{y \in D \setminus \{x\} | d(x, y) \leq kdist(x)\}$$

$$rd_k(x, y) = \max\{d(x, y), kdist(y)\}$$

$$lrd_k(x) = \frac{|N_k(x)|}{\sum_{y \in N_k(x)} rd_k(x, y)}.$$

これらの式を用いて，$LOF(x)$ は以下で定義される．

$$LOF(x) = \frac{1}{|N_k(x)|} \sum_{y \in N_k(x)} \frac{lrd_k(y)}{lrd_k(x)}$$

4.2 One Class SVM

One Class SVM は ν–SVM[1] を利用した外れ値検出の手法である．すべてのデータは $+1$ のクラスに属し，原点のみが -1 のクラスに属するとして，ν–SVM を使って 2 つのクラスを分離する超平面を求める．その結果，-1 のクラス側に属するデータを外れ値とする．

目的関数の式は以下である．

$$\min_{w,b,\xi,\rho} \frac{1}{2} w^T w - \rho + \frac{1}{\nu N} \sum_{i=1}^{N} \xi_i$$

subject to

$$w^T \phi(x_i) \geq \rho - \xi_i$$

$$\xi_i \geq 0 \quad (i = 1, 2, \cdots, N).$$

上記を以下の双対問題に変換して超平面を求める．

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha$$

subject to

$$0 \leq \alpha_i \leq \frac{1}{\nu N}$$

$$\sum_{i=1}^{N} \alpha_i = 1.$$

5 実験

一般の巨大コーパスとして新聞記事 3 年分 (毎日新聞 '93〜'95) を利用する．まずここから「災害」と「死者」というキーワードが含まれる文書を取り出し，それを文書セット 1 (文書数 329) とした．同様に,「女優」と「結婚」が含まれ文書セット 2 (文書数 163),「会社」と「倒産」が含まれ文書セット 3 (文書数 602) の合計 3 つの文書セットを作成した．

次に上記の 3 つの文書とは関連なさそうな文書を適当に 5 つ選び，それを外れ値文書セット 1 とした．また上記の 3 つの文書とは関連ないスポーツに関する文書を 5 つ選びそれを外れ値文書セット 2 とした．

文書セットと外れ値文書セットの組み合わせは 6 通りあるが，それぞれに対して，その外れ値文書セットを文書セット内に混ぜて，作られた文書セットから 3 つの手法 (LOF, One Class SVM 及び本手法) を用いて外れ値文書検出を行った．

LOF による実験結果を表 2 に示す．示した数値は 5 つの外れ値文書の外れ値の順位である．括弧の数値はそれら順位の平均である．

表 2: LOF での実験結果

文書セット	外れ値文書セット 1	外れ値文書セット 2
災害・死者	62,124,179,303,331 (200)	181,226,244,256,264 (234)
女優・結婚	27,28,35,135,161 (77)	5,25,30,49,60 (34)
会社・倒産	142,208,215,492,509 (313)	250,282,346,509,517 (381)

One Class SVM による実験結果を表 3 に示す．

表 3: One Class SVM での実験結果

文書セット	外れ値文書セット 1		外れ値文書セット 2	
	抽出数	正解数	抽出数	正解数
災害・死者	25	1	23	0
女優・結婚	46	1	50	1
会社・倒産	31	0	16	0

そして本手法による実験結果を表 4 に示す．

本手法と LOF は外れ値の度合いの順位を出力するので，単純に表 2 と表 4 を比較することで，手法のパフォーマンスを比較できる．明らかに本手法の方が優れいていることを確認できる．

One Class SVM は外れ値の度合いの順位を出力せずに，直接，外れ値を出力するので，単純には本手法

表 4: 本手法による実験結果

文書セット	外れ値文書セット 1	外れ値文書セット 2
災害・死者	2,6,8,16,48 (16)	1,2,3,5,7 (4)
女優・結婚	2,3,4,13,22 (9)	1,4,5,7,44 (12)
会社・倒産	3,17,18,65,122 (45)	2,10,12,28,43 (19)

と比較はできない．そこで本手法については表 4 で示した外れ値の平均の順位までを外れ値として検出したという仮定で F 値を出して比較した．その結果が表 5 である．なお，F 値は次式によって計算した．p は抽出した文書のうち，正解した文書の割合 (=精度)，r は外れ値文書のうち，抽出された文書の割合 (=再現率) である．

$$F = \frac{2 * p * r}{p + r}$$

表 5: One Class SVM (OCS) と本手法の比較

文書セット	外れ値文書セット 1		外れ値文書セット 2	
	本手法	OCS	本手法	OCS
災害・死者	0.381	0.067	0.667	—
女優・結婚	0.429	0.039	0.471	0.036
会社・倒産	0.120	—	0.250	—

表 5 より，本手法は One Class SVM よりも明らかに優れいていることが確認できる．

また外れ値文書セット 2 に対する検出結果を見ると，LOF も One Class SVM も外れ値文書セット 1 と同様に検出が困難になっているが，本手法の場合は逆に外れ値文書セット 2 の方がパフォーマンスがよく，外れ値文書の数に影響を受けないことがわかる．

6 考察

実験では，本手法は LOF や One Class SVM よりも良い結果が出せた．特に LOF や One Class SVM では検出率が悪くなると言われていた，同種類の外れ値文書に対しても良い結果となった．用意した外れ値文書よりも順位の高かったものを見ると，一週間の出来事の簡単なまとめなど，他の話題が多く混じっている記事であった．元から関連度の低い記事が混じっていたため，用意した非関連文書より高い順位になったと思われる．

改良すべき点が多くある．実験では一般の巨大コーパスとして新聞記事 3 年分を使用したが，この量が十分といえるかは検証する必要がある．また，この手法は逆トピックワードとして設定する単語数によって，結果が大きく変わる．設定する単語数があまりに少ないと，逆トピックワードが 1 つも含まれない文書が出てきてしまう．適切な単語数を決める方法が必要である．実験では一般の巨大コーパスからみても出現数があまりにも少ない単語はあらかじめ削除して行ったが，この少ないと判断する出現数にも適切な決定法が必要となる．

7 おわりに

本論文では，ある事柄に関連した文書セットを対象とした，外れ値文書の検出手法を提案した．外れ値文書の検出を行う場合，一般的にはデータマイニング分野の外れ値検出手法を用いる．しかし，この手法はある同類の外れ値文書がある程度の数出現した場合，検出に失敗しやすくなるという問題があった．そこで Bekkerman が提案した関連文書検出の手法を応用し，外れ値文書の検出を試みた．

実験では一般の巨大コーパスとして新聞記事 3 年分を利用した．そこからあるキーワードに関連した文書セットを作成し，そこから人為的に混ぜた外れ値文書の検出を試みた．結果は LOF と One Class SVM での実験結果と比較し，有効性を示した．

参考文献

[1] B. Schölkopf and J. C. Platt and J. Shawe-Taylor and A. J. Smola and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, Vol. 13, No. 7, pp. 1443–1471, 2001.

[2] Ron Bekkerman and Koby Crammer. One-class clustering in the text domain. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*, pp. 41–50, 2008.

[3] Markus M. Breuning and Hans-Peter Kriegel and Raymond T. Ng and Jörg Sander. LOF: Identifying Density-Based Local Outliers. In *ACM SIGMOD 2000*, pp. 93–104, 2000.

[4] 山西健司. データマイニングによる異常検知. 共立出版, 2009.