ARG WI2 Dec.14, 2013

様々なデータ圧縮手法を用いた tweets の話題分類の精度比較

王駿琪　佐藤栄一　澤勢一史　前川廣太郎　延原肇

筑波大学大学院システム情報研究科 知能機能システム専攻

概要　本稿では、データ圧縮手法に基づく tweets の話題分類の枠組みおいて、種々の圧縮手法の比較検討を行い、適切な手法の模索を行うとともに、エントロピー法を用いた手法の分類性能の評価を行う。

キーワード　話題分類　データ圧縮　エントロピー

1 はじめに

ソーシャルメディアの代表であるTwitterは、手軽に「今」を知る「今」を知らせることができるという特徴で、驚異的なスピードで成長を遂げている。一方で、やりとりされる情報の流れが速いため、興味のある情報を見落とす可能性が高く、これを解決するためにTwitterを対象とした情報推薦や情報検索などの研究が盛んに行われている。

Twitterの話題分類において、投稿(tweet)は「新語が多い」、「文法的誤りが多い」という特徴があるため、従来の形態素解析及び、bag-of-words[1]表現による機械学習では対応が難しく、データ圧縮を用いた情報類似度による分類手法が提案されている [2]。しかし、データ圧縮の手法は数多く存在するため、Twitterの話題分類に適切な圧縮手法を明らかにする必要がある。そこで、本稿では、deflate, gzip, bzip2, snappyとlz4の5種類の圧縮手法を用いてtweet分類を行う。また、圧縮の代わりにエントロピーを用いた場合についても比較する。

2 提案手法

データ圧縮とは、データが持つ冗長性を排除することで、データのサイズを小さくすることである。2つのデータを連結して圧縮する際、2つのデータの類似度が高いほど、冗長な部分が多くなり、圧縮ときのサイズが小さくなる。本稿で提案する手法は、西田らの手法[2]に基づき、指定した文字列(キーワード、ハッシュタグなど)が含まれるtweetのテキストを時間順に連結したものを話題モデルA、それ以外のtweetのテキストを時間順に連結したものを比較モデルBと定義する。それから、データの圧縮データ量について、本稿はBenedettoらの手法[3]に基づき以下のように計算する。

$$C_A(x) = Z(A + x) - Z(A) \qquad (1)$$

$$C_B(x) = Z(B + x) - Z(B) \qquad (2)$$

数式(1)(2)の中で、$Z(A + x)$はモデルAと入力tweetxの連結したデータの圧縮後のサイズを表し、$C_A(x)$はxとモデルAとの非類似度を表す。そして、tweetxに対する分類スコアは以下で定義する。

$$f(x) = \frac{C_A(x) + \gamma}{C_B(x) + \gamma} \qquad (3)$$

ここで、$f(x)$が分類閾値θより小さく、かつ、$C_A(x) < C_B(x)$のとき、tweet xはモデルBよりモデルAに類似していると判断できる。ここで、できる限り情報量の多いtweetを優先的に精度よく分類することを考慮したスムージングパラメータγを導入する。また、情報のエントロピー(平均情報量)の計算を行う。これにより、情報の圧縮限界が分かるため、各圧縮手法の性能を比較するには重要な指標として利用できる。エントロピーは数式(4)で示す。

$$H(X) = -\sum_{i=1}^{M} p_i log_2 p_i \qquad (4)$$

ここで、Mは情報に出現した文字の種類数であり、p_iは文字iの出現確率である。

3 評価実験

各圧縮手法の分類性能を評価するために、本研究では5分割交差検定を用いる。分類閾値θを変化させ、各手法の分類の再現率と適合率を示す。

3.1 実験データ

実験として、人気のハッシュタグを話題に設定、Streaming API を利用し、2013 年 11 月 2 日のラジオ人気番組「アニソンアカデミー」の放送時間に発信された tweet 合計 39426 件を収集した。収集した tweet のうち、番組のハッシュタグ“#aniaca”を付けていた tweet を話題モデル、ハッシュタグをつけていない tweet を比較モデルとした。ただし、tweet からリンクやハッシュタグをすべて除き、テキストだけに注目した。また、ユーザが検索を行うとき極端に短い tweet はユーザにとって有用ではない可能性が高いため、15 文字以下の tweet をすべて除取り除いて

いる。実験に用いた tweet データセットの件数を Table1 に示した。

各データセットはランダムに 5 分割し、1 つをテストデータ、残り 4 つを学習データとした 2 クラス分類実験を 5 回繰り返して各圧縮手法の分類性能を比較した（5 分割交差検定）。

Table 1 Tweets データセット

Tweets 種類	#aniaca	その他
Tweets 数	126	33306

3.2 実験結果

図 1 は、同じテキストデータへの適用に関して、lz4 や snappy に比べ bzip2, gzip および deflate が圧縮率の観点で優れていることを示している。また、エントロピー法の結果も同時に示している。さらに、話題分類の実験で、分類閾値θを変化させたときの各圧縮手法の適合率と再現率を図 2 に示した。分類精度について、deflate が高く、lz4 と比べて snappy のほうがやや高いことが分かった。図 3 により、実験で同じデータを処理した 5 種類の圧縮方法では bzip2 が一番時間がかかる。逆に lz4 と snappy の処理スピードが速い。結論として、tweets の話題分類において、五つの圧縮手法では、deflate が lz4 と snappy より適切であり、エントロピーによる話題分類の精度が悪いと判断できる。また、bzip2 について更なる検証が必要である。

4 おわりに

本稿では、5 種類の圧縮手法を用いて tweet の話題分類を行い、またエントロピーを用いた分類も比較した。評価実験の結果、bzip2 は適合率において、gzip を上回る性能を持つが、実行速度の点で劣る。また、圧縮が非常に高速な snappy や lz4 は、精度の点で他の圧縮手法に劣ることが評価実験により分かった。今後の展望として、1. 実験データを増やし、より正確に各圧縮手法の分類性能を把握する、2. 精度とスピードのバランスのとれた圧縮手法で tweet の話題分類を行うこと、3. 今回用いた 0 次経験エントロピーだけでなく、より良い精度を期待できる k 次経験エントロピーによる分類を行うこと、4. データ圧縮による分類は言語によらないため、日本語以外の言語にも適用すること、が挙げられる。

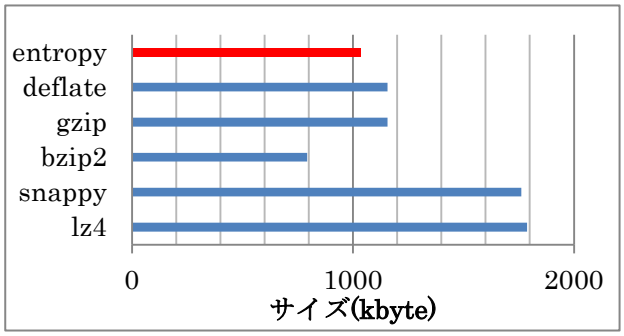

図1 圧縮したサイズとエントロピーの比較図

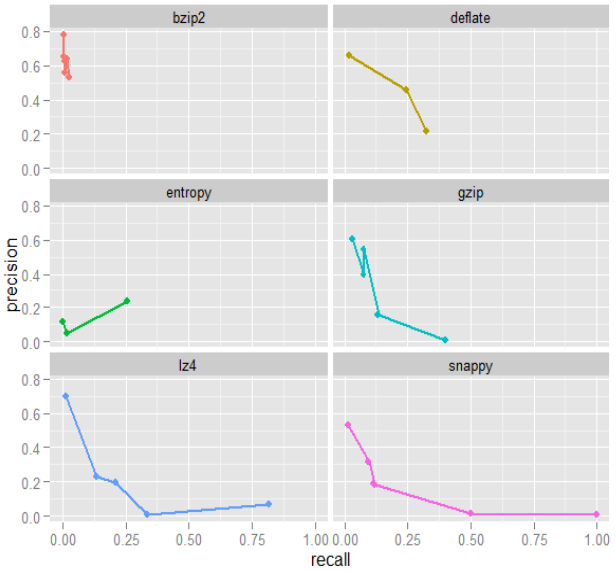

図2 適合率と再現率の結果図

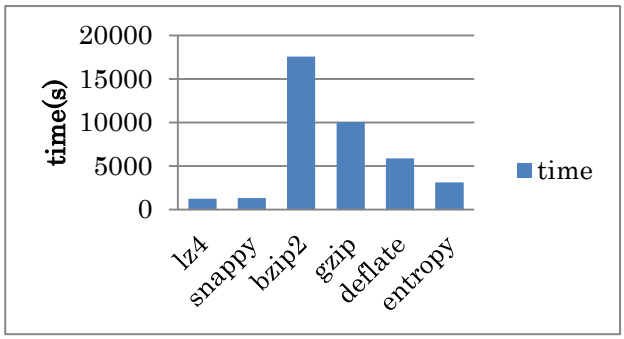

図3 各圧縮法の執行時間

参考文献

[1] Sivic, Josef (April, 2009): Efficient visual search of videos cast as text retrieval. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 31, NO. 4. IEEE. pp. 591–605.

[2] 西田京介，坂野遼平，藤村考ほか：データ圧縮によるTwitter のツイート話題分類，日本データベース学会論文誌10(1), 1-6,2011-06-00.

[3] D. Benedetto, E. Caglioti, and V. Loreto: Language trees and zipping, Physical Review Letters,vol.88, no.4, 2002.