機械学習分散処理フレームワークによる概念辞書構築

尾脇拓朗† 福元伸也† 赤木康宏† 河合由起子†† 川崎 洋†
† 鹿児島大学 †† 京都産業大学

1. はじめに

近年, Web 上のデータから単語の意味的関係を示す概念辞書の構築に関する研究が行われている. 我々は, 概念辞書を用いてユーザの潜在的な検索要求に対応することのできる検索システムの開発を目指している. 本研究では, 機械学習分散処理フレームワークを用いた概念辞書の構築について報告する.

2. 分散並列型の機械学習について

世界中で日々, 膨大なデータが発信されている. これらのビッグデータと呼ばれる大量のデータは, データベースのように構造化されておらず, その量の多さのため, これまでは分析されてこなかった. しかしながら, Hadoop などの分散並列処理システムがオープンソースとして提供されるようになり, ビッグデータの分析に関する研究がにわかに注目をあびるようになってきている. この大量のデータを分析するための1つの手段に機械学習フレームワーク Jubatus がある[1]. Jubatus は, データを抽象化して特徴ベクトルに変換し分析を行う.

3. 提案システム

本研究で用いるシステムの概要を図 1 に示す. まず, Web 上の情報を取得するためのクローラーに Apache Nutch を使用して, ニュース記事を収集する. 収集された記事は, 今回は, Yahoo API を利用して形態素解析を行い, その中から名詞の単語だけを抽出した. 抽出された単語を学習器

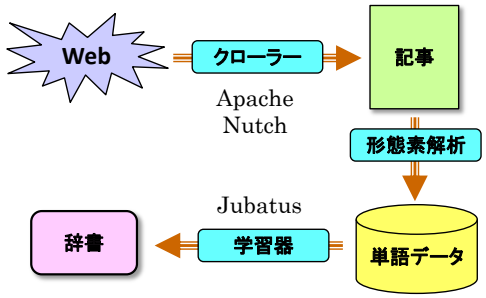

図 1. システム概要

Jubatus に与えることにより, クラスタリングされた分類結果が得られる. この結果を用いて辞書を構築する.

4. 実験

実験では, 辞書カテゴリの識別率の評価を行った. 単語データは, ニュース記事において「経済」, 「スポーツ」, 「コンピュータ」の 3 つのカテゴリから記事を収集し, 形態素解析で名詞だけを抽出し, それぞれのカテゴリに属する単語を 600 個集めた. その中から, トレーニングデータとして, 500 個, テストデータとして, 100 個を使用した. この単語データを Jubatus に与え, テストデータで識別したところ, 表 1 のような結果が得られた. 識別結果は, テストデータが, どのカテゴリの属しているかで判別され, 識別率 86%という結果が得られた.

表1 データ数と識別率

トレーニングデータ数	1500 (=500×3)
テストデータ数	300 (=100×3)
識別率	86 %

5. おわりに

本研究では, 並列分散処理に対応した機械学習器を用いて, ニュース記事に含まれる単語の識別を行った. その結果, 学習データがあまり多くないにもかかわらず, 比較的良い識別率が得られた. 今後, データ数を増やし, 辞書の構築を行っていく予定である.

6. 謝辞

本研究の一部は, 内閣府 NEXT(LR030), 総務省 SCOPE(101710002,102107001), JST A-STEP (AS242Z02958H) および JSPS 科研費(24500120, 24780248)の助成を受けて実施された.

参考文献

[1] 小田 哲, 中山心太, 上西康太, 木下真吾: “Jubatus：Big Data のリアルタイム処理を可能にする分散処理技術, ”信学技報 IN, vol. 111, no. 409, pp. 35-40, 2012.