

第1四半期報告会 ツアーマイニング

氏名：片倉 翔太+タイメンバー

報告年月日：2017年7月12日

1. はじめに

私たちは、SNS 主に twitter を用いて不特定多数の人の旅行記録を抽出し、それらをラベリングしデータベースに保管しておき、後にユーザがキーワードを打ち込むとそのキーワードに関連する旅行記録を出力し、ユーザの興味や嗜好に合わせた旅行プランの提案を目標とし、タイからの留学生も含めたメンバーで研究を行なっている。

またメジャーな場所だけでなく、マイナーであるが特定のユーザの興味をそそるような旅行プランの提案もできるように日々研究を行なっている。

2. 準備

- Twitter API
- google map API
- DB
- Mecab・kuromoji
- Web アプリケーション

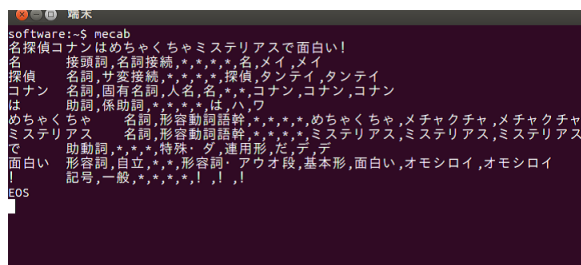


図 1: Mecab

3. 目標

API によるデータの収集をする際には、twitter における各ユーザのツイートはタイムリミット(収集可能期間)がある。そこで我々は膨大なデータをラベリングして分類してオフラインの DB に保管しておく。また web アプリケーションでユーザがキーワードを打ち込むとそのキーワードに対応するラベリングされた過去の旅行記録を出力する。

4. 研究成果

・ツアーレコードを取得し、それらツアーレコードの緯度と経度を用いて google map に表示できるようにした。また、それらのデータから個人情報が

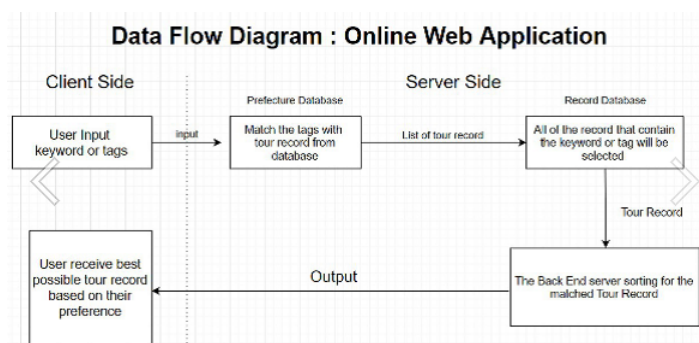


図 2: DFD

わからないようにクリーニングし、県名で表示されるようにした。

- ・twitter のユーザの分析を行うために形態素解析の Mecab を使えるようにした。
- ・収集したデータ(ツイート)を手動でラベリングし、web アプリケーションを作成した。
- ・web アプリケーションではキーワードを打つと、キーワードとマッチする旅行記録のツイートを表示するようにした。

5. 研究課題

- ・例として「常磐公園」とサーチをかけて検索すると、北海道の常磐公園に関するツイートが含まれていたりノイズとされる情報が多く含まれていたりするため、ノイズの除去などが必要である。
- ・API を用いた時に収集できるデータには期限がある。
- ・莫大なツイートの中で有益な旅行記録をピンポイントで収集すること。
- ・データ収集やラベリング作業における手動と自動の使用範囲が不明瞭である。自動類推システムの構築。
- ・位置情報を含むツイートが少なく、なおかつ位置情報を含むツイートに関しても都道府県によって人口面でのツイートの偏りが生じている。

6. 研究計画

- ・位置情報付きのツイートは非常に少ない問題に対応するために、位置情報に頼る必要のない言語処理での判別を行い、「地名+行った!」というツイート等から自動で旅行記録を取得できるように

する。(8-10 月)

- ・ユーザの趣味をラベリングする。旅行記録をツイートしたユーザのいいねを読み込ませ名詞のみを抽出し、抽出された名詞から趣味や関連キーワードを類推する。それらユーザの趣味や関連キーワードと実際の旅行の相関性を分析する。(8-10 月)

- ・Mecab の拡張。「名探偵コナン」といったキーワードを名・探偵・コナンと区切るのではなく「名探偵コナン」で抽出できるようにする。(8 月)

7. 発表計画

- ・研究会 11/16,17 @日立中央研究所 (CAS, IPSJ-AL) 参加予定

参考文献

[1] 高橋一希, “マイクロブログユーザの地域による嗜好の違いの分析,” DEIM Forum, B3-3, 2016.

[2] 難波 英嗣, “観光情報の自動編纂,” 知能と情報 (日本知能情報ファジィ学会誌), vol.26, no.1, pp.9-15, 2014.