7K-01

Bag of Words と skip-gram 併用によるレビュー・店舗間類似度評価とそれに基づく店舗推薦

山本真史　山崎俊彦　相澤清晴

東京大学

1　はじめに

Yelp [1] などに代表されるレビューサイトの普及により，個人でも気軽にレビューをネット上に投稿することが可能となった．それに伴い，大量に発信された情報の中から個人が必要とする情報を検索することが困難になっている．個人で有用な情報を検索することが困難になるに従い，大量の情報からユーザーが求める情報を提示する情報推薦の重要性が増している．本研究では，ユーザーが過去に書いたレビューを利用し，ユーザーに対して適切な店舗推薦を行うことを目的とする．

2　関連研究

類似ユーザーや類似アイテムの情報を用いてユーザーに推薦を行う協調フィルタリングは広く推薦システムにおいて広く使われており，さまざまな研究が行われている．例えば B. Sarwar ら [2] は映画推薦を目的として，アイテムベースの協調フィルタリングを用いて Mean Absolute Error が 0.72 の精度でユーザーの映画に対する評価予測を行っている．協調フィルタリング以外にもユーザーの行動を予測して推薦を行うシステムの研究が行われており，W. Zhang と J. Wang [3] は地方の新規イベントに対するユーザーの反応を予測している．

3　提案手法

たとえば東京大学周辺にある山手ラーメンは東京大学の学生もよく利用するラーメン屋であるが，そのような現地の人になじみ深い店舗をユーザーがあまり詳しくない地域でも探すことができるようになることを想定する．そこでユーザーが過去に書いたレビューを利用し，実際にそのようなレビューが書けると考えられる店舗の推薦を目的とする．本研究では解く問題を既存レビューの入力に対してそれがどの店舗に対するものかを推定する問題として設定する．その第一歩として，クエリレビューと店舗に対して書かれたレビュー集合を一つのレビューとした場合の店舗との類似度を計算することで入力レビューがどの店舗について書かれたかを推定する．

3.1　使用データ

Yelp はデータの一部を Yelp's Academic Dataset [4] として公開している．本研究では Yelp's Academic Dataset から，181,680 人のユニークユーザーが 16,465 店舗に対して書いた 679,382 レビューを用いた．この 679,382 レビューから最新の 1,000 レビューを抽出したのち，抽出した 1,000 レビューについてどの店舗に対して書かれたかの推定を行った．

3.2　評価店舗予測

本研究では Bag-of-Words (BoW) と word2vec [5] という手法で提案されている skip-gram (skip-gram) を組み合わせた評価店舗予測実験を行い，比較対象として BoW ならびに skip-gram のそれぞれ単体を用いた場合の評価店舗予測実験を行った．

はじめに提案手法である skip-gram と BoW を組み合わせた手法について説明する．あらかじめ wikipedia から skip-gram の辞書を作成し，単語をクラスタリングすることで単語とクラスタの対応付けを行う．skip-gram

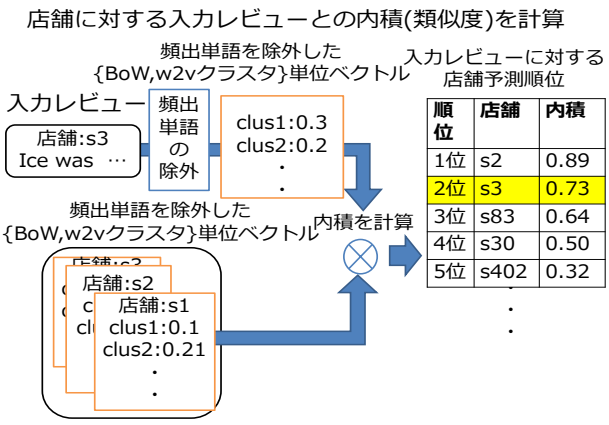

図. 1　評価店舗予測概要

Shop Recommendation Based on Review-Shop Similarity using Bag of Words and Skip-gram

Masafumi Yamamoto, Toshihiko YAMASAKI, Kiyoharu AIZAWA

The University of Tokyo

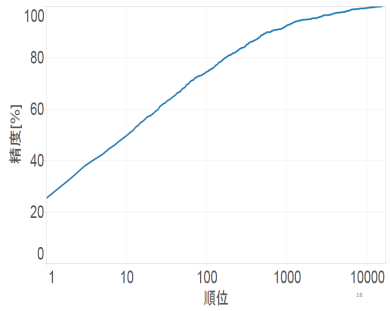

図. 2　BoW に基づく予測結果

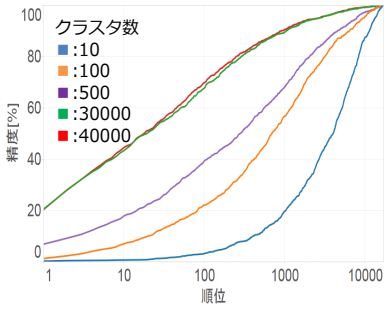

図. 3　skip-gram に基づく予測結果

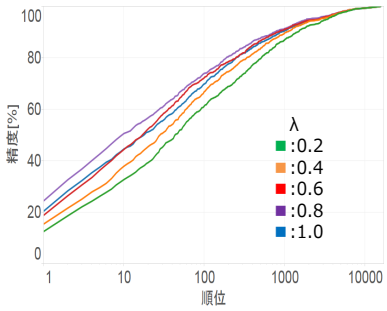

図. 4　提案手法による予測結果

の辞書に登場する単語に対しては出現回数で重み付けした skip-gram クラスタ単位ベクトルを作成し，skip-gram の辞書に登場しない単語に対しては出現回数で重み付けした BoW 単位ベクトルを作成する．skip-gram クラスタ単位ベクトル間内積 Sim_w，BoW 単位ベクトル間内積 Sim_b を計算し，

$$(\text{レビュー・店舗間類似度}) = \lambda Sim_b + (1-\lambda) Sim_w \quad (1)$$

で記述される各類似度の線形和に基づく店舗類似度ランキングを λ を作成し，評価店舗予測を行う．なお，ここでは skip-gram のクラスタ数は予備実験の結果から 40000 を利用した．

次に比較対象の BoW，skip-gram のそれぞれに基づく評価店舗予測についてその概要を図 1 に示す．BoW に基づく評価店舗予測では，クエリレビューと店舗に対して頻出単語を除外したのちに出現回数で重み付けけした BoW 単位ベクトルを作成する．これらのベクトル間内積に基づく店舗類似度ランキングを作成し，評価店舗予測を行う．skip-gram に基づく評価店舗予測では，クエリレビューと店舗に対して頻出単語を除外したのちに skip-gram の辞書内単語に対して出現回数で重み付けした skip-gram クラスタ単位ベクトルを作成する．これらのベクトル間内積に基づく店舗類似度ランキングを作成し，評価店舗予測を行う．skip-gram のクラスタ数は 10, 100, 500, 30000, 40000 で実験を行った．

また，除外した頻出単語については今回は予備実験から全店舗中 4 割以上の店舗で利用された単語を頻出単語として類似度計算をする際に除外した．

4　実験

BoW に基づく実験結果を図 2 に，skip-gram に基づく実験結果を図 3 に，skip-gram のクラスタ数 40000 の時に BoW と skip-gram を組み合わせた場合の実験結果を図 4 に示す．グラフの横軸がランキング順位，縦軸がその順位までに正解店舗を予測できたレビューの割合を表す．上位 10 位までに正解店舗を BoW 利用時に 49.8%，skip-gram 利用時ではクラスタ数を増やすに従い精度が向上し，クラスタ数 40000 の時に 44.6%，BoW と skip-gram を組み合わせることで λ=0.8 とした場合に 50.6% の精度で予測できている．BoW と skip-gram を組み合わせることにより提案手法では BoW と skip-gram のそれぞれを単独で用いた場合よりも正解店舗を精度良く予測できていることがわかる．

5　まとめ

ユーザーにとって有益な情報を得ることが困難な現代において，ユーザーに対して適切な店舗を推薦することが本研究の目的である．その第一段階として，今回は Yelp's Academic Dataset を用いてクエリレビュー・店舗間類似度を計算することにより評価店舗予測実験を行った．上位 10 位までに BoW と skip-gram を組み合わせることで 50.6% の精度で正解店舗を予測することができた．今後は類似度計算方法の改良や，類似度計算時に用いる単語の条件づけなどによる正解店舗予測精度の向上を目指す．

6　謝辞

本研究の一部は科学研究費助成事業 (26700008)，および Microsoft IJARC core 10 の支援プログラムを受けて行われた．

参考文献

[1] `http://www.yelp.com`.
[2] B. Sarwar et al. Item-based collaborative filtering recommendation algorithms. in WWW, 2001.
[3] W. Zhang and J. Wang. A collective bayesian poisson factorization model for cold-start local event recommendation. in KDD, 2015.
[4] `https://www.yelp.com/academic_dataset`.
[5] T. et al Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.