第３回勉強会　レジュメ

B5EB1106　酒井洋輔

タイトル：Beyond the Stars :Improving Rating Predictions using Review Text Content

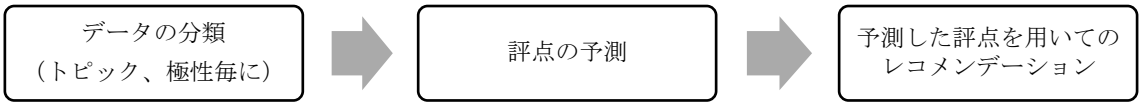
著者：Gayatree Ganu, Noēmie Elhadad , Amelie Marian

掲載：12th International Workshop on the Web and Databases ·WebDB 2009
(Providence, Rhode Island, USA)

【研究概要】

ある商品 a に評点（星の数など）をつけた消費者 i は、別の商品 b にどれほどの評点をつけるだろうという予測を立てる。そして予測評点が高い商品を消費者 i に推薦するというのが、レコメンデーションの仕組みの一つである。

この研究では、評点だけを考慮していた従来の方法に、テキスト情報を加えることで、予測精度が向上するということを示した。



1. Introduction

インターネットのますますの普及により、WEB 上には大量の商品のレビューテキストデータが蓄積している。しかし余りにも膨大であるため、目当ての情報を探し当てることもまた難しくなりつつある。特に消費者が商品についてあまり知識を持っていない場合、消費者から商品についての情報を得ることは困難であるため、**レコメンデーション※１**が必要となってくる。

※１ レコメンデーションには

①協調フィルタリング…「商品 A を購入した人は商品 B を購入する傾向があるため、A を購入した人に B を薦める」。という手法。閲覧履歴、購入履歴を使用。

②コンテンツベースフィルタリング…商品の属性があらかじめグループ化されており、ユーザーにはそのグループ内からレコメンドする手法

③　①と②のハイブリッド　　という種類がある

2．データの分類と分析

【用いたデータ】

Citysearch という口コミサイトの、レストランのレビューデータ 50000 件（2006　地域はニューヨーク）

レストランの数は 5531 軒、32284 人の識別されたユーザーを含む

【文の分類（Manual）】

①6 つのカテゴリー（Food, Service, Price, Ambience 雰囲気,　Anecdotes 伝聞性, Miscellaneous その他 ）

②4 つの極性（Positive, Negative, Neutral, Conflict 良いことと悪いことどっちも書いてある ）

に、3400 文を手作業で分類する。

450 文を、3 人の異なる判定者によって分類してみて、**K 係数※2** でその信頼性を測る

Ancdotes 以外は概ね信頼性は確認された。

※2 複数の判定者の一致度を測る指標

【文の分類（Automatic）】

上記のテストデータを用いて、SVM(サポートベクトルマシン)で分類する。

方法は **K-cross validation※3** を用いる。

その結果を Accuracy(正解率),Precision(適合率),Recall(再現率)で評価する

※3　K 分割交差確認法…K 個にデータを分割して、その内の一つをトレーニングデータとして使って予測を行い、残りの K－1 個のデータで結果を評価する。を K 回繰り返す方法

【レビーテキストの分析】

以下の結果が得られた

・文全体の 56%が positive な文であった。対して、18%が negative な文であった

・32%の文が Food について、17%が Service について、10%が Ambiance(雰囲気)、6.5%が price について書かれていた

・上記のトピック別の割合は、料理のジャンル（中華、イタリアン…）によって異なり、例えばフレンチとイタリアンは Service についての言及が多かった

・Price が上がると、値段についての positive な意見が減り、negative な意見が増えた

Figure1 参照

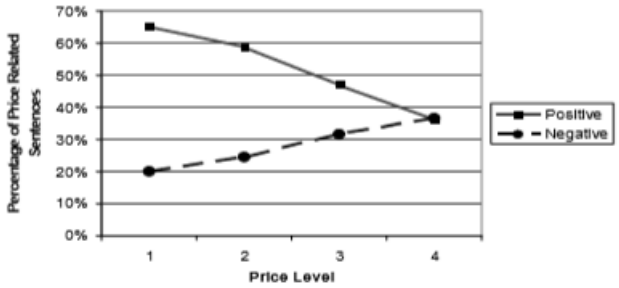


Figure 1: Impact of price level on perception.

【文の極性と評点の関係の分析】

Positive な文、negative な文を含む割合と、評点の相関係数を計算すると、

評点～positive：0.45

評点～negative: －0.48

また、評点が 4～5 の高評価なレビューの 71%が positive な文を含んでおり、評点が 1～２の低評価なレビューの 78%が negative な文を含んでいた

以上のが、テキストデータを評点の予測、レコメンデーションに組み入れた方が良いのではないかというモチベーションとなる

３．評点の予測

仮説：**従来使われている予測評点は、他のレビューの平均点である。テキストデータから、従来の方法よりも精度の高い予測評点が作れるのではないか**

評価方法：約 260 件のレビューを二種類抽出（Test Ⅰ、Test Ⅱ）

Test Ⅰ→12 件以上のレストランのレビューを書いているユーザーのレビュー

Test Ⅱ→5 件以上のレストランのレビューを書いているユーザーのレビュー

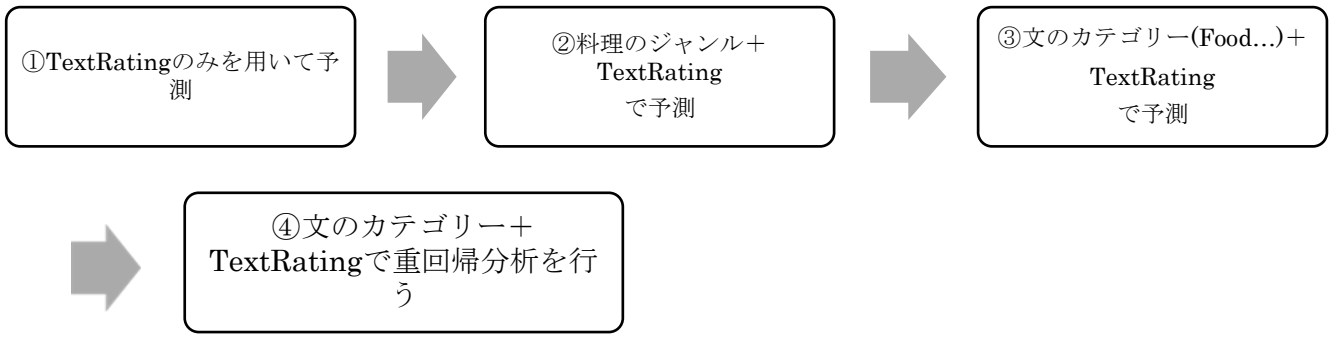
MSE (Mean Square Error)を指標として採用

テキストデータを予測評点に変換：　$$TextRating = \left[\frac{P}{P+N} * 4\right] + 1$$

P、N→それぞれ、あるレビューに含まれる positive な文、negative な文の数

※Neutral と Conflict は有益な情報が含まれていないとして考慮しない

以下の４つの条件で予測を行う



【①の条件】

結果は Table2 の通りとなった。比較対象となる、評点のみを用いた予測評点は、あるレストランについての他の評点の平均値を用いた。

Predicting Star Ratings	TEST I	TEST II
Star rating	1.217	1.295
Sentiment-based text rating	1.098	1.27
Predicting Text Sentiment Ratings	**TEST I**	**TEST II**
Star rating	1.430	1.342
Sentiment-based text rating	1.277	1.374

Table 2: Prediction MSE using the restaurant average for prediction.

ウィルコクスン検定（マンホイットニーの U 検定）※4 を用いて、Star rating と Sentiment-based text rating の結果に有意な差が出ているかどうかを確認する

その結果、Test Ⅰについては p=0..02 で有意な差が確認できたが、Test Ⅱでは p=0.12 で有意な差が確認できなかった。

表の上部は Star Rating を推定したもの、下部は Text Sentiment Ratings を推定したものとなっている。

Text Sentiment Ratings において結果が悪化したのは、文には多様性があるためであると考えられる。

※４ウィルコクスン検定（マンホイットニーの U 検定…２つの集団の変数を比較したい場合で、正規分布に従っていない時に用いる

【②の条件】

①では、StarRating ,TextRating 共にある一軒のレストランの評点の平均点を用いた。

②では、メタデータ（料理のジャンル、ロケーション、価格帯）が同じようなレストラン全体の平均点を用いる。

結果は Table3 の通りである。

StarRating と TextRating に①ほどの大きな差が見られなかった。推察される理由は本文にはフワッとしか書かれていない

Predicting Star Ratings	TEST I	TEST II
Star rating	1.030	1.117
Sentiment-based text rating	1.051	1.135
Predicting Text Sentiment Ratings	**TEST I**	**TEST II**
Star rating	1.245	1.233
Sentiment-based text rating	1.275	1.199

Table 3: Prediction MSE using cuisine average for prediction.

【③の条件】

小仮説：**カテゴリーの中には、予測に有用なもの、そうでないものがあるのではないか**

結果は Table4 の通りとなった。

表の上部は、それぞれ該当のカテゴリーに属する文の極性のみ用いて評点を予測した。

①と比べても、あまり良い結果は得られていないが、その中でも Food は比較的良い結果が得られている。

料理に関する情報は、評点を予測する上で有用なのではないかと推測される。

表の下部は、**表の上部であまり良い結果が得られなかったカテゴリーを除くと、①の条件よりも予測精度が向上することを示している。**

Predicting Star Ratings	TEST I	TEST II
Food	1.215	1.308
Price	1.377	1.424
Service	1.531	1.623
Ambience	1.427	1.559
Anecdotes	1.57	1.676
Miscellaneous	1.221	1.436
All but Food	1.130	1.281
All but Price	**1.096**	1.279
All but Service	**1.096**	**1.269**
All but Ambience	1.115	**1.264**
All but Anecdotes	**1.096**	**1.254**
All but Miscellaneous	1.181	1.352

Table 4: Prediction MSE using the restaurant average for prediction, only considering some categories for the text ratings.

カテゴリー毎に予測に対する有用度の差があることが、③の条件よりわかった。
④の条件では、重回帰分析を行うことで、**カテゴリーごとの weight を明らかにすることが目的である。**

従属変数：StarRating
独立変数：カテゴリー、極性で分類した文が占める割合

$$\text{StarRating} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots$$
$$X_1 = Food\ Positive,\ X_2 = Food\ Negative\ \ X_3 = Service\ Positive \ldots$$

結果は Table5 と Table6 の通りとなった。
Table6 を見ると、③の条件で得られた結果通り、Food の weight が大きくなっている。
Regression では結果が評点の範囲（１～５）を超えてしまうため、Table5 では標準化された結果が、StarRating と比較されている（Table2 と同じもの）

Predicting Star Ratings	**TEST I**	**TEST II**
Star rating	1.217	1.295
Regression-based text rating (scaled)	1.089	1.231
Predicting Regression-based Text Ratings (scaled)	**TEST I**	**TEST II**
Star rating	2.680	2.461
Regression-based text rating (scaled)	2.593	2.414
Predicting Regression-based Text Ratings (raw)	**TEST I**	**TEST II**
Regression-based text rating (raw)	0.702	0.742

Table 5: Prediction MSE using the restaurant average for prediction, two-sentiment regression.

Regression Weights	Positive	Negative
Food	4.86	1.53
Price	1.67	1.59
Service	2.61	0.51
Ambience	2.35	2.43
Anecdotes	3.65	2.02
Miscellaneous	5.17	2.27

Table 6: Two-sentiment regression weights.

４.Personalized Recommendation

個人の嗜好に合わせたレコメンドを提案する

KNN 法（K 近傍法）を使って、ユーザーをクラスタリング（?）する。

あるユーザーがあるレストランについてレビューしていない時、その代替値として

平均値を使う。

本研究の新規性は、この代替値にテキストデータを用いることによって予測精度を上げたことにある。

5. 今後の展望

・個人の興味、レビュースタイルに基づいたユーザーのグループ分けをしていきたい

・ユーザーが自分の興味のあるカテゴリー、極性に絞ったレビューを探すのを援助したい

参考サイト

Silver Egg Technology

https://www.silveregg.co.jp/archives/blog/49