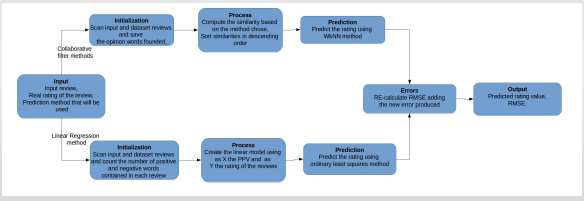
Four different ways to predict reviews' rating through text analysis

Davide Lissoni Daniele Dellagiacoma
University of Trento

Problem & Solution

In the last few years, progress in digital data acquisition and storage technology has led in the growth of huge datasets. For this reason, **data mining** has recently emerged as an interdisciplinary subfield of computer science. Data mining has the goal of extracting as much information as possible from a dataset and transform it into an understandable and useful structure for further use. The automatic or semi-automatic analysis of huge quantities of data aims to discover interesting patterns such as groups of data records, unusual records and dependencies. These patterns may be used in further analysis or in machine learning and predictive analytics. Another relevant technique related to data mining is **text mining**. It has the purpose of extracting meaningful information from a text. Nowadays, more and more users can freely express their opinions about products or services on Internet. These reviews can be very useful both to other users for making informed decisions and to companies for upgrading their services. Moreover, the amount of reviews grows rapidly and understanding their role in e-commerce has become an important topic both for academics and practitioners.

We thought about an algorithm which could be used to predict the rating of an input **text review**, based on all others rating reviews in the dataset. We compare the input review with each other review in the dataset to determine which are the most similar to the input review.
We determine three different similarities using a set of **positive and negative opinion words**.
Finally, we use the computed similarities and **linear regression** for predicting the ratings of the input review.



Four different ways

Cosine similarity
It measures the cosine of the angle between two vectors. Its outcome is bounded between 0 and 1. It is commonly used in text mining to determine the similarity between two texts. We counted the number of occurrence of each opinion word in the texts.

$$CosSim(x,y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

Jaccard similarity
It is commonly used to comparing the and the diversity of sample sets. It is defined as the size of intersection divided by the size of the union of the sample sets.

$$JaccSim(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Pearson correlation coefficient
It calculates the correlation between two or two vector, returning a value between 1 and -1 where is positive correlation, is no correlation and -1 is negative correlation.

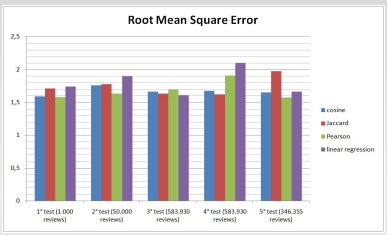
$$PearCorr(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

Linear regression
We used the rating of the reviews as dependent variable y and the Positive Predict Value of each text review as independent variable x.

$$PPV = \frac{n^\circ\ positive\ words}{n^\circ\ positive\ words + n^\circ\ negative\ words}$$

Result

We reported 60 different tests on 5 different Amazon dataset. We achieved the best results using the cosine similarity collaborative filter methods that has had a **RMSE** equal to 1,66980667, predicting in total 13 right values.



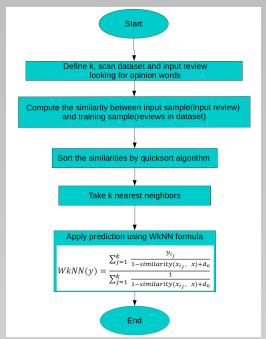
The results mean that, the similarity between two reviews expressed in order to predict their rating based on the opinion word that they contains, is more precise if it will be calculated on the occurrences of the equal words that are written in the two reviews (cosine similarity) and not on a correlation between them (Pearson correlation) or on how many equal words are present in the two reviews (Jaccard similarity). Our linear regression system is in general the method that produced the worst results between those that we developed

Implementation

We provide a reviews' rating prediction method, that, given as a review and a reviews dataset try to recommend the rating of the input review, comparing this with each dataset review. We do prediction using two different recommender systems: Collaborative filter methods and Linear Regression.

Collaborative filters

1. Look for reviews which use the same opinion words with the input review (the review whom the prediction is for);
2. Use the ratings from those like-minded reviews found in step 1 to calculate a prediction for the input review;
3. a value by a WkNN algorithm, using the similarity method input-choice.



Linear Regression

1. Compute on each dataset review the PPV by using the opinion words as "important" words;
2. Create a Simple linear model, using as independent variables the PPV calculated, while as dependent variables the rating of the review;
3. Calculate the PPV on the input review and estimate its rating using ordinary least squares method on the linear model just built.