Unsupervised Rating Prediction based on Local and Global Semantic Models

Adrian Boteanu, Sonia Chernova
Worcester Polytechnic Institute
100 Institute Road Worcester, MA 01609
aboteanu@wpi.edu, soniac@wpi.edu

Abstract

Current recommendation engines attempt to answer the same question: given a user with some activity in the system, which is the next entity, be it a restaurant, a book or a movie, that the user should visit or buy next. The presumption is that the user would favorably review the item being recommended. The goal of our project is to predict how a user would rate an item he/she never rated, which is a generalization of the task recommendation engines perform. Previous work successfully employs machine learning techniques, particularly statistical methods. However, there are some outlier situations which are more difficult to predict, such as new users. In this paper we present a rating prediction approach targeted for entities for which little prior information exists in the database. We put forward and test a number of hypotheses, exploring recommendations based on nearest neighbor-like methods. We adapt existing common sense topic modeling methods to compute similarity measures between users and then use a relatively small set of key users to predict how the target user will rate a given business. We implemented and tested our system for recommending businesses using the Yelp Academic Dataset. We report initial results for topic-based rating predictions, which perform consistently across a broad range of parameters.

Introduction

One common method of making recommendations for users is based on identifying similar previous cases and using them for constructing relevant answers. The comparison criteria has significant influence on the output. In this paper, we explore the potential of global and local semantic models have in predicting user preferences. The main goal of this work is to estimate how a user would rate a business (a restaurant, for example) which he or she never visited before, based on the user's reviews of other establishments. Unlike sentiment analysis approaches such as (Qu, Ifrim, and Weikum 2010), we do not use the review text to predict it's rating; instead, we only use the users' review history.

This task presents a significant challenge: users with few reviews to date might not fit statistical models. For example, a frequency approach might limit suggestions within a particular cuisine because the user has a single review about a restaurant of that certain category.

For this project we used the Yelp's Academic Dataset, which contains 229,907 reviews, written by 43,873 users for 11,537 businesses from the same area. Our work focused on the *review*, *user* and *business* collections. In total, we used the following attributes:

- Review: business_id, user_id, stars, text;
- User: user_id;
- Business: business_id, stars, review_count

We purposely ignore the business category tags, assuming reviewers prefer to have a palette of options to chose from. This is motivated by the empirical observation that many words from each review are describing service quality and the reviewers' experience instead of category specific foods or beverages. We base our predictions for a given target user on other people that have previously reviewed some businesses in common with the target user – the *predictor users*. We give definitions and notation throughout the paper.

We extend the work of (Boteanu and Chernova 2013) in constructing topics using unstructured fact networks, such as ConceptNet (Havasi, Speer, and Alonso 2009). By using a concept graph, the approach leverages semantic similarity to associate words into topics, while being robust to syntactic errors. Using an adaptation of these methods we build global and local topic models to enhance rating accuracy, starting from the predictor users. The main contributions of this paper are the global and local methods of using topics to compare users for review similarity. Local topics are derived only from a single user's reviews. Thus, the method can scale to any number of users, since model complexity is a function of the number of reviews per user only.

The amount of data available per user varies greatly, from a single short review of a few sentences to hundreds of in-depth reviews. Modeling users via semantic topics allows us to have a more reliable model over statistical approaches for those users with little data. Furthermore, by using topic overlap measures which take into account semantic similarity between words, we allow for different expressions of the same concept, for example synonyms. The size of the topic model is dependent only on the vocabulary size, which tends to not expand significantly as the number of reviews per user increases, making our approach scalable for prolific users.

We present initial results supporting the assumption that topic models, especially local topics, are viable predictors of potential ratings. We then explore a machine learning approach of integrating multiple prediction methods into a classifier to predict ratings based on a single other user.

Related Work

Numerous recommendation systems focus on predicting which item, be it a film or restaurant, a user will enjoy based on their history and other users' history. The implicit result of any such system is that it selects items that the user might rate the highest. In this paper, we estimate how a user would rate an entity if he/she were to rate it. We only take into account the users' previous reviews and their relation to others users' ratings and reviews. Predicting a numeric rating instead of simply recommending the entity produces a recommendation gradient which is meaningful for the user, avoiding the dichotomy of other recommendation systems. This gradient could be used, for example, in conjunction with other costs such as travel distance.

Many topic modeling approaches focus on statistical NLP. The well known LSA (Landauer, Foltz, and Laham 1998) and other newer approaches use latent Bayesian models to uncover the underlying semantic associations between words based on their co-occurrence in sets of documents. While this approach has proven to be highly successful for large data collections, the common sense topic method we use in this paper scales from comparing pairs of single reviews to comparing large sets. Also, by having common sense semantics drive the topic formation, the same methods can be extended to any sparse data prediction scenario.

Very successful recommendations, in general, can be made using statistical data. For instance, (Li, Guo, and Zhao 2008) use statistical methods to filter tags that users are interested in from social media. Because our methods scale down, high confidence predictions can be made even for fairly new or less active reviewers. Also, (Linden, Smith, and York 2003) propose using as multiple sources of information about a users' behavior on the site as possible so that the lack explicit user reviews is compensated, similar in spirit with our motivation for using a decision tree to make predictions.

As noted by (Symeonidis, Nanopoulos, and Manolopoulos 2010), statistical methods are a simple but obtuse method of recommendation. The authors consider integrating tag, keyword and frequent item recommendation methods by using semantics to link each method. Our solution also uses semantics, but since no business type tags are used (although the application data set contains them), it is unsupervised.

(Ganu, Elhadad, and Marian 2009) use semantics and sentiment analysis to obtain a better understanding of the message a review is conveying, linking star ratings with user sentiment. Their work facilitates searching in a large number of reviews. We are estimating how a user might evaluate a business he/she never visited before, thus the target review cannot be analysed. As opposed to top down approaches such as ontology based recommendation frameworks (Yu et al. 2007), the local models we present in this paper can successfully predict ratings starting from only a single other review.

good wonderful love
ask call thank
home place house
steak burger food sweet bacon ...

Table 1: Examples from our global topic model

Topic Modeling

(Boteanu and Chernova 2013) propose a semantic topic modelling approach which uses an unstructured commonsense network, ConceptNet (Havasi, Speer, and Alonso 2009), to segment the vocabulary extracted from text into topics. The authors define topics as sets of words with no other associated metadata; the same word can simultaneously belong to multiple topics.

We adapted this model for our project. The previous work describes the topic creation as a two stage process: an initial topic aggregation followed by a refinement process. We found that for our purposes in this project, the overhead added by the topic refinement stage has little bearing on the prediction performance, so we use raw topics directly.

The original paper assumed that the entire vocabulary is available before the topic creation begins. We implemented the algorithm for a streaming architecture, in which words are added to topics as the corpus is processed. In the preprocessing stage, we remove stopwords and convert the remaining words to lemmas using the Stanford NLTK. The original paper presents an interest metric which is used to select words from the dialog that speakers a interested in. We do not perform this dialog-specific processing since the reviews in our corpus are structured very differently.

By including every non-stop word in the topic creation process, our topic vocabulary is a much larger proportion of the text. Furthermore, because our application has a much broader context than the story used by (Boteanu and Chernova 2013), the potential vocabulary also broader. Thus, we limit the number of words in the vocabulary when constructing the global topic set to the first two thousand distinct lemmas. Even so, there are 3245 topics in the global model. Some examples are shown in Table 1.

Notation

The goal of our project is to predict the rating which a given user assigns to a given business, using only the users' other reviews; we do not analyze the review text associated with the rating we are trying to predict. In this section we define the various sets of users and businesses we operate with in our algorithms, followed by a compact notation description.

We define *target rating* as the rating which we are estimating. This is included in the review written by the *target user*, u_t, to the *target business*, b_t. From the total number of businesses reviewed by the target user, we select a subset, which does not include the target business. We call this the set of *predictor businesses*, $BPred(u_t)$. Similarly, we define *predictor users*, $UPred(u_t)$ as the set of users that reviewed the predictor businesses, excluding the target user.

It is very uncommon to find more than 30 reviewers who wrote about the exact two businesses, as it also is very uncommon to have more than six businesses in common between two reviewers. We set a upper limit on the number of the $BPred(u_t)$ and $UPred(u_t)$ in order to reduce the bias towards very popular businesses which are reviewed by prolific reviewers.

We denote individual users and businesses as u_i and b_j, respectively. Sets of users that reviewed a set of businesses are written as $U(B)$; similarly, $B(U)$ stands for the set of business IDs reviewed by users in U; By r_{u_i,b_j} we denote the text of the review written by u_i for business b_j. Analogously, s_{u_i,b_j} is u_i's rating in stars for b_j; we denote its predicted value by $s^*_{u_i,b_j}$. In the following section we introduce further notation as needed.

Proposed Methods

Our first hypothesis is that the predictor users' ratings for the target business are a better estimate of the target rating than the target business average. Our second hypothesis is that weighing in the semantic similarity between users lowers prediction error, as opposed to weighing each user equally. In this section we present three methods for predicting the the target rating from a small subset users. The first hypothesis is tested by all three methods, on top of which the global and local topic models are both motivated of the second hypothesis.

We use each method in this section to compute the weighted average rating of $s(u_i, u_t)$, for all $u_i \in UPred(u_t)$. Each calculates the weight for each predictor user, $w(u_i)$, differently. The final prediction value for the target rating is

$$s^*_{u_t,b_t} = \Sigma w(u_i) * s(u_i, b_t), u_i \in UPred(u_t) \qquad (1)$$

Predictor Users

This method explores our first hypothesis, that users in $BPred(u_t)$ offer better indication for $s(u_t, b_t)$. $BPred(u_t)$ also serves as a subset of users small enough to make semantic reasoning feasible. For this method, all $w(u_i)$ are equal. Thus, we compute the predicted rating as the arithmetic mean of the ratings users in $BPred(u_t)$ gave to b_t, with all weights in Equation 1 equal to $1/UPred(u_t)$.

We consider this prediction method our baseline, which we aim to improve using topic similarity. This prediction, as all presented methods, ignores some causal relations, such as the order in which the businesses in the predictor set were visited with respect to the target business. Such considerations are beyond the scope of this project, which assumes that businesses are fairly stable over time and users rate consistently. For example, (Potamias 2012) points out that, in reality, businesses initial rating diminish slightly over time under certain conditions; however, the difference found by is much smaller than our baseline accuracy.

Semantic Similarity

Building semantic topic models is a costly process. However, we can infer various measures of similarity between users by applying such models. Starting from our second hypothesis, we investigate two approaches to obtain semantic similarity ratings when dealing with a very large number of users:

- A *global topic model*, which is used to summarize and compare two sets of reviews.
- Multiple *local topic models*, one per user, used for pairwise comparisons of users.

We define $sim((u_i, B_k), (u_j, B_l))$ as the similarity rating for a pair of users, comparing topics derived from reviews written by u_i for businesses in B_k to topics derived from reviews written by u_j to businesses in B_l.

Global Topic Models Our first approach is to build corpus-wide topic models. This is motivated by the assumption that topics are immutable, thus the same topics will convey the same meaning for all users. In this view, topics are intrinsic to the corpus, or more generally, to the language. Unlike building user-specific topics, this method allows to have a common unit of measure when comparing any two users. For a very large set of reviews, $R(\{u_i\}, B)$, we take the vocabulary of all reviews written by u_i to form a single set of words in lemma form. To keep the model to a reasonable size, we limit the size of this vocabulary to two thousand words, taken in order. At this scale, selecting the most frequent words did not provide any advantage.

We construct topics starting from this vocabulary using the method described in the Topic Modeling section. Then, when predicting $s(u_t, b_t)$, we generate an activation vector for each user in $UPred(u_t)$ and u_t, in which a topic is marked as active if at least one word in the user's vocabulary is present in that topic. The result is an activation vector of $True$ or $False$ values for to each topic in the global model for each user in $UPred(u_i) \cup \{u_t\}$ for $BPred(u_t)$.

Before computing the similarity between each user from $UPred(u_t)$ and u_t, we filter out irrelevant topics. We compute the Pearson correlation coefficient between topic activation and the users' ratings. We only consider topics for which each the correlation coefficient is significant at level $\alpha = 0.05$, or if the topic is active for all users in $UPred(u_t)$.

We define the *similarity* between two such vectors as the proportion of matching activation topics between two activation vectors. We use the set of global topics to compute two measures of similarity:

- $sim_p((u_i, BPred(u_t)), (u_t, BPred(u_t))), u_i \in UPred(u_t)$: How similar were the predictor users and the target users in writing about the predictor businesses;
- $sim_c((u_i, Bpred(u_t)), (u_i, b_t))), u_i \in UPred(u_t)$: How consistent are the predictor reviews with the review about the target business.

We normalize these weight vectors, then normalize their pairwise product, $sim_p(u_i) * sim_c(u_i)$ to obtain the final weights, $w(u_i)$, which are used in Equation 1.

Local Topic Models Local topics are constructed individually for each user, starting from the predictor reviews. Each predictor user, along with the target user, has a separate local topic model. The set of local topics is computed for each

user $u_k \in UPred \cup \{u_t\}\}$, using all the reviews that u_k wrote about businesses $b_k \in BPred(u_t)$.

Since two reviewers have reviewed few businesses in common, it is tractable to build topic models for each by using their entire predictor vocabulary. The main motivation of this approach is that no arbitrary bounds on the vocabulary or the number of reviews are set, since they are expected to be computationally feasible.

In this model, we compute the similarity between users as the average pairwise overlap of topics from each users' model. We define overlap as the proportion of the intersection of the two topic, plus the average ConceptNet Divisi similarity distance between the words in their difference. The normalized similarity ratings are used as weights for the predictor ratings to estimate the target rating. We note that we also experimented with alternatively using the Wordnet path similarity metric (Pedersen, Patwardhan, and Michelizzi 2004) between words, which led to worse results.

For comparison, we also consider a bag of words model, in which each users' vocabulary is used as a single most general topic, similar to (Qu, Ifrim, and Weikum 2010).

Implementation and Experimental Results

We integrated the above methods into a benchmarking framework to systematically evaluate their individual performance. There are two main modes of operation:

- The first, which we call *aggregated rating*, creates predictions from normalized weights outputted by each method described in the previous section. The final ratings are compared with the target rating for accuracy.
- The second, which we call *individual predictions*, is to take into consideration the predictive power of each predictor user, by feeding each set of individual weights and the predictor users' rating for b_t into a decision tree.

Aggregated Prediction Error Results

We evaluate the prediction performance of the following methods:

- Random: a randomly selected integer from 1 to 5;
- ZeroR: the most common rating in the dataset is 4;
- Business Average: the target business average rating;
- Baseline: average rating of $UPred(u_t)$ for b_t;
- Bag of Words: to explore the benefit of constructing topics from the review vocabulary, we include a prediction method which uses the entire vocabulary of $R(\{u_i\}, BPred(u_t)$ as the most general topic;
- Local Topic Model;
- Global Topic Model.

Table 2 shows the mean, standard deviation and median for the error and absolute error of our prediction relative to the target rating. In calculating each prediction error, we rounded the estimated ratings to the nearest half star in order to have the same granularity as the business average ratings.

With the exception of randomly assigning ratings, there is little difference between the other prediction methods. Due to the distribution of ratings between reviews, even the ZeroR predictions are not significantly off on average. This is the result of having, out of the total number of reviews, over 30% of reviews 4 or 5 stars, each. Low ratings are uncommon, thus the ZeroR predictor is fairly reliable.

We explore the relation between prediction accuracy and high level statistics about each business. We computed monotone regression splines using *proc transreg* from SAS 9.2 to express the absolute error as function of the number of reviews b_t has, the number of users in $UPred(u_t)$ and the average rating of b_t, respectively. These plots are shown in Figure 1. These figures reveal some interesting information which is not apparent when judging only the mean values.

The number of predictor users is strongly connected to the absolute error of all of our methods. Relative to each methods' performance, having lots of predictor users dramatically lowers the prediction error. The same trend can be also observed in the second graph: as the number of reviews of target business increases, the absolute error decreases. However, it is important to note that the errors resulted from the ZeroR rule and average business predictions also decrease, at the same rate, as our models. This shows that the size of the population which we use to predict is not the only factor contributing to the decrease in error. The third graph hints to the fact that because ratings of 4 or 5 stars are so common, simply predicting a high rating is fairly successful.

The regression curves of the third graph also show that poorly rated businesses are evaluated very differently between individuals. Because the business average also shows a higher error rate for poor ratings, it indicates that there is a mix of inconsistent, high and low, ratings.

Overall, we observe that the local topic models generally outperform both the global topic models and the bag of words comparison, especially in the case of small numbers of predictor users. These initial results motivate further research in more reliable topic comparison methods, which would more strongly emphasize user similarities and difference. However, the business average remains the strongest overall predictor, due to relative low variance of ratings.

Individual Prediction Decision Tree Classification

Our second method of evaluation focuses on how accurately an individual reviewer from $UPred(u_t)$ can predict the target rating. Instead of weighting the ratings $s(u_i, b_t)$, we build a classifier which take in some parameters for each user and the target rating. This is motivated by highly similar results of all our methods, despite observing very different weights assigned to each predictor user.

Each instance has the following attributes:

- counts: number of reviews for b_t, number of predictor users
- non-semantic: business average rating, baseline weight;
- semantic: bag of words weight, local topic model weight, global topic model weight;
- prediction class: $s(u_t, b_t)$, five classes.

Starting from 3932 instances, we constructed seven J4.8 decision trees in Weka to predict the target ratings from only some attribute subsets shown in Table 3. Using 10-fold cross validation, we obtained the prediction accuracy shown in the same table.

Index	Prediction	Mean Err	Std Dev Err	Median Err	Abs Mean Err	Std Dev Abs Err	Median Abs Err
1	Random	-0.852	1.842	-1	1.620	1.223	1
2	ZeroR	0.165	1.186	0	0.866	0.827	1
3	Business Average	-0.002	1.104	0	0.847	0.708	0.5
4	Baseline	-0.038	1.312	0	0.990	0.862	1
5	Bag of words	-0.124	1.325	0	1.058	0.977	1
6	Local topic model	-0.420	1.759	0	1.293	1.265	1
7	Global topic model	-0.054	1.342	0	1.002	0.894	1

Table 2: Aggregated Prediction errors for each method, rounded to three decimal points

Index	Counts	Non-Semantic	Semantic	Acc
1	•			91.94%
2		•		43.44%
3			•	92.74%
4	•	•		90.46%
5	•		•	90.46%
6		•	•	54.74%
7	•	•	•	92.5%

Table 3: Prediction accuracy of J4.8 decision trees incorporating the marked sets of features.

From visually inspecting the tree structures, we noticed some interesting patterns. The top level in each tree that includes non-semantic attributes is the average rating of b_t. Another very high ranking metric is the number of reviews a business has. This points to the fact that if a business is popular enough to receive many (good) reviews, then it has many satisfied customers who write positive reviews.

However, the pruned decision tree constructed using all available parameters (last in Table 3) contains 563 nodes and 282 leaves, with a maximum depth of 26. Local and global similarity ratings are used starting from depths of eight and nine, respectively, confirming that these parameters have some predictive power. This can be also observed by the slight increase in prediction performance when these parameters are included next to the non-semantic attributes. 2.

The business average is placed at the root of all trees which include the attribute, with split between less than four stars and greater or equal than four stars. Overall, the decision tree analysis further confirms that it is the poorly rated businesses that are more difficult to predict for. These have a mix of satisfied and unsatisfied customers. Otherwise, for a very popular location, there is little doubt that a customer will rate it positively. In the latter case, the business average is a better predictor since the ratings have low variance.

To further underline the difference in prediction accuracy between high ratings and low ratings, we include graphical visualizations of the confusion matrix results for two target ratings, 2 and 4, in Figure

Conclusion and Future Work

In this paper we presented our initial results on using global and local semantic models to predict how a user would rate a business if he/she were to rate it. We implement and test our models on the Yelp Academic Dataset, a large collection of business reviews from the Phoenix, Arizona, area. Our system adapts existing topic models, scaling them for large scale data. We believe that our approach, with the local topic-based similarity in particular, has the potential to be a highly effective method of predicting individual users' assessments of new and not very well know businesses, for which little review data is available.

We present two models for predicting ratings based on these similarity models. The aggregated predictions rely on using a single user similarity estimation to output the predicted value. In our experiments, we show that each method has strengths and weaknesses depending on the particular characteristics of each target user and target business. Our second prediction model sets on tying together all available prediction metrics and generating classifications based on evidence provided by a single other user, which shares some background with the target user. Again, this approach lends itself especially for lesser known businesses.

In the case of highly popular business, we believe there is little improvement to be made in the case of the average user. These locations are popular as a result of their customers' satisfaction, so reviews tend to be favourable. However, even for such businesses, we see great potential in our decision tree prediction method to identify those customers that have opinions different from the consensus.

For our future work, we plan on focusing on integrating multiple knowledge bases together for evaluating topic overlap, in order to better predict user similarity. As noted in the paper, we experimented with using both Wordnet and ConceptNet word similarity, but only separately since the two have very different design philosophies. Furthermore, we plan to give our prediction model temporal awareness, such that trends in the way a business is regarded are taken into account during the prediction process.

Acknowledgements

This work was supported by National Science Foundation award number IIS-1117584.

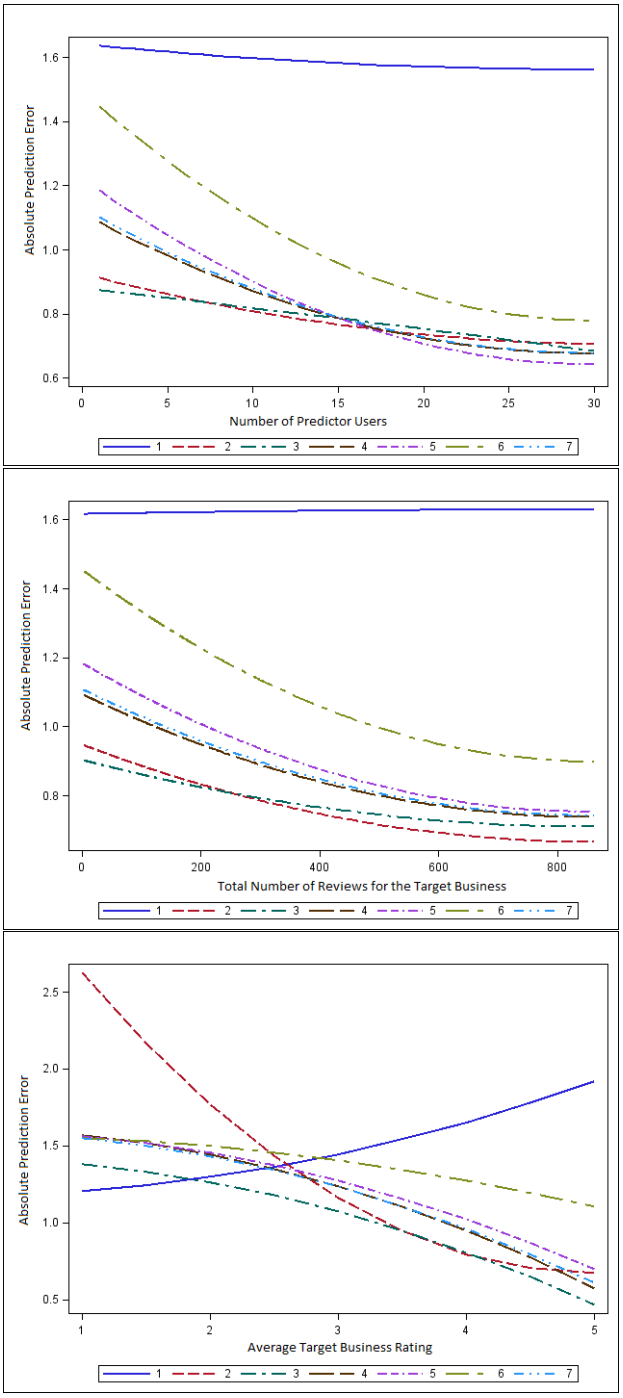

Figure 1: Monotone regression spline for the absolute prediction error as function of the number of similar users (top), number of reviews b_t has (middle), average rating of b_t (bottom). The prediction method indices correspond to Table 2.

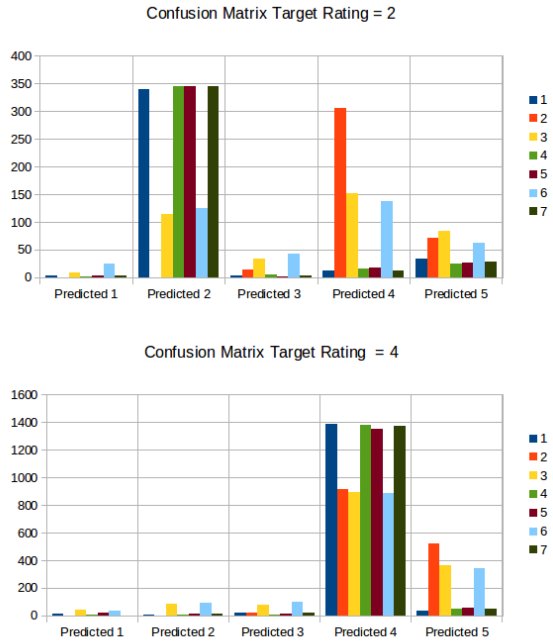

Figure 2: Confusion matrix results for ratings 2 and 4, with indices for decision trees from Table 3.

References

Boteanu, A., and Chernova, S. 2013. Modeling discussion topics in interactions with a tablet reading primer. In *Proceedings of the International Conference on Intelligent User Interfaces*.

Ganu, G.; Elhadad, N.; and Marian, A. 2009. Beyond the stars: Improving rating predictions using review text content. In *12th International Workshop on the Web and Databases*.

Havasi, C.; Speer, R.; and Alonso, J. 2009. *ConceptNet: A lexical resource for common sense knowledge*, volume 5. John Benjamins Publishing Company.

Landauer, T. K.; Foltz, P. W.; and Laham, D. 1998. An introduction to latent semantic analysis. *Discourse processes* 25(2-3):259–284.

Li, X.; Guo, L.; and Zhao, Y. E. 2008. Tag-based social interest discovery. In *Proceedings of the 17th international conference on World Wide Web*, 675–684. ACM.

Linden, G.; Smith, B.; and York, J. 2003. Amazon.com recommendations: item-to-item collaborative filtering. *Internet Computing, IEEE* 7(1):76–80.

Pedersen, T.; Patwardhan, S.; and Michelizzi, J. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, 38–41. Association for Computational Linguistics.

Potamias, M. 2012. The warm-start bias of yelp ratings. *arXiv preprint arXiv:1202.5713*.

Qu, L.; Ifrim, G.; and Weikum, G. 2010. The bag-of-opinions method for review rating prediction from sparse

text patterns. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 913–921. Association for Computational Linguistics.

Symeonidis, P.; Nanopoulos, A.; and Manolopoulos, Y. 2010. A unified framework for providing recommendations in social tagging systems based on ternary semantic analysis. *Knowledge and Data Engineering, IEEE Transactions on* 22(2):179–192.

Yu, Z.; Nakamura, Y.; Jang, S.; Kajita, S.; and Mase, K. 2007. Ontology-based semantic recommendation for context-aware e-learning. In *Ubiquitous Intelligence and Computing*. Springer. 898–907.