Recommendation Using Textual Opinions

Claudiu-Cristian Musat, Yizhong Liang, Boi Faltings
École Politechnique Fédérale de Lausanne
Switzerland
{firstname.lastname}@epfl.ch

Abstract

Many web sites collect reviews of products and services and use them provide rankings of their quality. However, such rankings are not personalized. We investigate how the information in the reviews written by a particular user can be used to personalize the ranking she is shown. We propose a new technique, topic profile collaborative filtering, where we build user profiles from users' review texts and use these profiles to filter other review texts with the eyes of this user.

We verify on data from an actual review site that review texts and topic profiles indeed correlate with ratings, and show that topic profile collaborative filtering provides both a better mean average error when predicting ratings and a better approximation of user preference orders.

1 Introduction

E-commerce portals such as Tripadvisor, Expedia or Amazon collect numerous reviews of the products and services they market. These reviews are often used to rank items so that users can easily find the ones with the best rating. Today, this ranking is carried out without any personalization, so every user sees the same ranking. It would be nice to personalize this ranking so that users could be directed to hotels or products that fit their specific preferences.

Applying traditional collaborative filtering to such rating sites is difficult because we rarely find a sufficient overlap between products and services rated by different users, making it hard to judge user similarity based on similarity of review scores. We thus explore a different approach where we use the review texts as a basis for judging similarity: two users are similar if their review texts address the same topics. This allows comparing users even when they did not write reviews of the same hotel, and thus enables collaborative filtering in spite of the sparse data.

The underlying intuition is that a user's reviews reflect her personal interest: someone who is most interested in cleanliness, like the user in the example presented in Figure 1, will write about how clean a hotel is, but maybe not about its location. For this user, a ranking based on other reviews that also placed a lot of importance on cleanliness would be more meaningful than one that placed equal importance on every aspect. Moreover, it would provide an incentive for regular users to contribute reviews, as these will help provide better personalization for their own use.

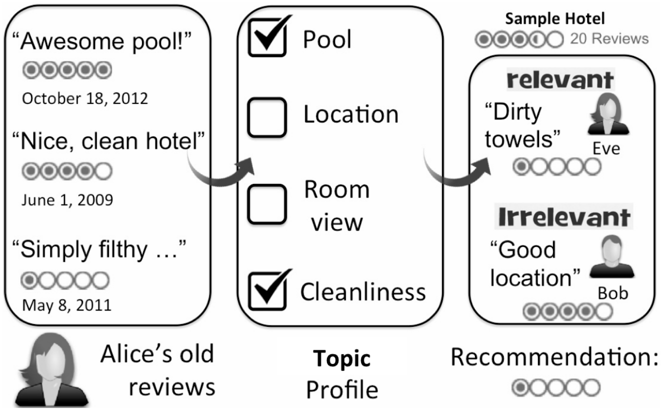

Figure 1: Topic Profile Collaborative Filtering.

To apply this intuition, we propose a new form of collaborative filtering, called *topic profile collaborative filtering - TPCF*, illustrated in Figure 1. Based on the topics a user writes about, we create an individual interest topic profile containing the topics that the user has expressed opinions about. For instance, in Figure 1 example, Alice has previously commented on the pools and the cleanliness of the venues she previously stayed at. We interpret this as a higher interest in these topics than others, like the hotel's location. We aggregate this information in her user profile. We then personalize the product rankings for each user, based on the reviews that are most relevant to her profile. Relevance is computed using the similarity between the topics commented upon in the reviews and the topics present in the profile. In the example, we will base our recommendation on the review that comments on the towels, rather than the one discussing the location.

We investigate the correlation of review texts and ratings using data from a popular hotel review site. Using techniques of automated opinion extraction, we first verify the basis of our approach by showing that opinions in review texts are indeed strongly correlated with a user's rating, but less correlated with other user's ratings. We further show that these opinions are in particular expressed on the topics in their topic profile.

We then show that our topic profile collaborative filtering produces a more accurate estimate of user ratings, using the standard measure of mean average error.

A more meaningful assessment of the impact on actual recommendations can be obtained by considering the ranking of items that corresponds to the predicted scores. Rankings are usually compared using measures such as Kendall's τ measure [Kendall, 1938] that are based on counting how many pairs of items are placed in opposite orders. As the sparsity of data does not allow us to compute this for all ranked items, we focus only on pairs where we are able to compute personalized scores and show that TPCF produces more accurate orderings than a non-personalized ranking.

The paper is structured as follows: the related work is outlined in Section 2 and the proposed method follows in Section 3. Section 3.1 details the extraction of interests and opinions from review text and Section 3.2 defines the topic profiles and the recommendation method that uses them. Section 4 shows the experimental results.

2 Related Work

Data sparsity is one of the main reasons personalizing product ranking is difficult. Because of it, products cannot be reliably linked to users [Kazienko and Kolodziejski, 2006]. A possible response is to increase the data available to the method. This often leads to hybrids between collaborative filtering[Sarwar *et al.*, 2001] and content based methods [Kazienko and Kolodziejski, 2006; Schein *et al.*, 2002; Levi *et al.*, 2012]. Social relations were also used to complement traditional data, either in the form of social tags [Zhang *et al.*, 2010] or social trust [Pitsilis and Knapskog, 2012].

Creating social tags or defining social trust is, however, a costly process. In general, all data acquisition above the normals workflow is intrusive and counterproductive [Middleton *et al.*, 2004], as users are reluctant to either rate items or answer questions [Sugiyama *et al.*, 2004]. [Pitsilis and Knapskog, 2012] also stress the importance of acquiring user relation data without extra effort from users. We believe this is also the case when we elicit the user's preferences with regard to a product or service's aspects. While other approaches use information like the browsing history [Middleton *et al.*, 2004] to define a user, we propose using the text of the user's previous writing as the relevant information source.

Most recommender systems do not use textual information to generate recommendations. This is probably due to the difficulties of getting machines to understand written natural language at a level sufficient to compete with simpler data sources, such as star reviews or buying patterns. Most methods rely on counting words, term frequencies or topic signatures[Balabanović and Shoham, 1997; Pazzani and Billsus, 1997; O'Mahony and Smyth, 2010]. A significant extension is the extraction of textual opinions. Several approaches have been proposed. Unigram models that extract the overall review polarity [Poirier *et al.*, 2010] have generated modest improvements. [Snyder and Barzilay, 2007] were among the first to observe that extracting overall opinions is not helpful, and relied on faceted opinions instead. They analyzed restaurant reviews to infer the author's sentiments regarding several aspects (e.g. food and service) Recently [Homoceanu *et al.*, 2011] have shown that faceted opinion mining extracts valuable information that helps cluster reviews and provide decision support to users. [Jakob *et al.*, 2009] also complement star ratings with opinions about predefined topics, and show marginal improvements.

A recent approach by [Levi *et al.*, 2012] is the most similar to our work. They propose recommendations using a combination of features, building a vocabulary for hotel aspects, extracting sentiment towards them and profiling users by using additional knowledge such as nationality or the purpose of the trip. The user's preferences are manually extracted, and the evaluation is based on questionnaires, which are difficult to apply on a large scale. Moreover they view the topics interesting to the users as only the context in which a recommendation is made. We believe the topics that users express interest in are the most important information available and model our recommendation accordingly. In this respect, we have a view similar to that of [Hariri *et al.*, 2011], who model the user's probable current needs, inferred from the trip type.

Personalization is the heart of recommendations, but evaluating its benefits is difficult. The traditional way of measuring the performance of recommender systems is to quantify to what extent they can predict the numeric rating a user gives to an item. The mean average error (MAE) or root mean squared error (RMSE) are the most widely used metrics [Shani and Gunawardana, 2009]. This is true for most systems that use the opinions within the free text of reviews [Faridani, 2011; Snyder and Barzilay, 2007], which use the star ratings as a benchmark for the opinion mining task. We believe this approach is not entirely consistent with the idea that the text contains information which is unavailable by examining only the numeric rating. If the opinions expressed within the text component are fully contained in the star ratings, there is little use of the additional component. More recent work [Levi *et al.*, 2012] used a user satisfaction evaluation for their recommendation method. However asking hundreds of people to voluntarily provide feedback is not easily scalable. We propose two additional evaluation methods that do not require users to depart from their normal workflow and do not rely solely on predicting the absolute value of the rating.

3 Proposed Method

3.1 Interpreting Review Content

Relevant Aspects

We rely on the intuition that, in their reviews, people leave opinions about multiple independent aspects. The aspects can be grouped together into classes, or topics. Let V be the used vocabulary, and $\mathcal{P}(V)$ its power set. In an initial preprocessing phase we identify a vocabulary subset corresponding to all the nouns, $V_n \in V$, with a power set $\mathcal{P}(V_n)$. The intuition is that people generally offer opinions about nouns.

We first find the *relevant aspects* discussed $A \subset V_n$. We then group them into a set of topics $Z = \{z_1..z_m\}$ where $m \in \mathbb{N}$ is predefined and Z represents the set of all possible topics. There are multiple competing definitions for aspects, from manually created ones, to frequency or relation based techniques and topic modeling.

The accuracy of the recommendations depends heavily on the aspect modeling choice. We created two different topic sets. The first set was obtained using Latent Dirichlet Allocation [Blei *et al.*, 2003] on a corpus consisting of 68049 reviews. For the second set, we used a frequency based technique. The nouns in the same review corpus were ordered by the number of attached opinions and we selected the ones with the highest opinion counts. We present the manner in which an opinion is attached to a noun below. The remaining nouns were then manually grouped into 18 topics. For each noun, we selected its Wordnet [Miller, 1995] synsets which were likely to appear in the review data and added those synonyms to their topics.

To compare the two topic sets, we created a human computation game. The players of this game were asked to judge whether a pair of statements, taken from two different reviews, actually discussed the same aspect. The pairs were constructed using nouns from the same topic. When the pairs were constructed with LDA topics, the majority of the pairs were judged to be incorrect. When they were constructed with the opinion frequency approach, the majority of the pairs were marked as correct. We thus used the latter topics.

Faceted Opinion Extraction

Next, we find the opinions present in the reviews with regard to each aspect, known as faceted opinion extraction. We identified all the words that are in a direct syntactic relation with the important aspects discussed. Then, for each identified aspect, we select the related words that represent an opinion and aggregate the result into a subjectivity value for the aspect.

For phrases contained in reviews from a collection $p \subset r \in R$, where $p \in \mathcal{P}(V)$, we construct a set of syntactic relations [de Marneffe *et al.*, 2006] that contain the phrase's nouns. A relation is defined on a pair of words and takes predefined values corresponding to known relation types, $\mathcal{R}$, $\rho : V \times V \to \mathcal{R}$. To reduce the amount of noise, we limit the relations that we consider to the ones typically attached to opinionated utterances, $\mathcal{R}' \subset \mathcal{R}$. For instance, we consider relation types like *nsubj* relevant, while ones like *prep* irrelevant. For each noun of the analyzed phrase, we extract the other phrase words with which it has a meaningful relation: $\rho_n : V_n(p) \to V(p), n \mapsto \rho_n(n) = \{w|\rho(w,n) \in \mathcal{R}'\}$. For instance, in the phrase *"The room was too small and the carpets dirty"*, the noun *room* is related to the adverb *small*, but not to *dirty*

For a given phrase p, we first determine the relation set for each present aspect instance, $\rho_n(a), a \in A \cap p$. We then perform a word level subjectivity analysis in which we use known polarized words from OpinionFinder [Wilson *et al.*, 2005]. Each word $w \in V$ has a corresponding polarity value $pol(w) \in \mathbb{N}$. In this case the values range from -2 to 2, with negative values for words like *bad* or *horrendous*, positive values for ones such as *wonderful* or *excellent* and zero for words not present in the OpinionFinder list.

The subjectivity score of an *aspect instance* a within a phrase $p \in \mathcal{P}(V)$ is defined as the sum of the polarities of the words in $\rho_n(a)$: $subj : A \times \mathcal{P}(V) \to \mathbb{N}$:

$$subj(a,p) = \sum_{w \in \rho_n(a)} pol(w) \tag{1}$$

Based on it, we define the aggregate subjectivity value of an aspect *within a review* or review set R as the sum of the subjectivity scores of the aspect's occurrences within that text or text collection: $subj : A \times \mathcal{P}(V) \to \mathbb{N}$,

$$subj(a,R) = \sum_{p \in R} subj(a,p) \tag{2}$$

We extend the subjectivity definition to a topic as the sum of the subjectivity towards the component aspects, $subj : Z \times \mathcal{P}(V) \to \mathbb{N}$,

$$subj(z,R) = \sum_{a \in z} subj(a,R) \tag{3}$$

and the overall subjectivity as the sum of the subjectivities over all the topics in the set: $subj : \mathcal{P}(V) \to \mathbb{N}$,

$$subj(R) = \sum_{z \in Z} subj(z,R) \tag{4}$$

We define the number of opinions uttered about an aspect in a phrase in a similar manner, without considering the polarity sign, and the opinion count with respect to a given rating or topic is defined as above, $count : A \times \mathcal{P}(V) \to \mathbb{N}$:

$$count(a,p) = \sum_{w \in \rho(a)} |sign(pol(w))| \tag{5}$$

3.2 Topic Profile Collaborative Filtering

As shown in Figure 1, the novel recommendation method we propose uses the reviews the user has written previously to create an interest profile which is then used to select suitable reviews.

Let $R_{p,i}$ be the reviews that user i has previously written. The preference profile of user i is then modeled as set of all topics whose opinion count in $R_{p,i}$ exceeds a certain topic significance threshold ts:

$$Z_i = \{z_i | count(z_i, R_{p,i}) > ts\} \tag{6}$$

In our experiments, we set $ts = 0$ as the sparsity of the data would otherwise leave us with many very small profiles; we expect that in a more active rating site this could be set to a higher value.

For each product A, let $r_{j,A} \in R_A$, $j \in 1..|R_A|$ be its reviews and $sr_{j,A}$ be the associated ratings. For user i with interest profile Z_i, we define a weight that is proportional to the number of topics in Z_i addressed in the review $r_{j,A}$:

$$Z_{i,r_{j,A}} = \{z \in Z_i | count(z, r_{j,A}) > 0\} \tag{7}$$

Rather than ranking products by their average review score, in TPCF we use the following weighted average:

$$TPScore_{i,A} = \frac{\sum_{j<|R_A|, j \neq i} sr_{j,A} \cdot |Z_{i,r_{j,A}}|}{\sum_{j<|R_A|, j \neq i} |Z_{i,r_{j,A}}|} \tag{8}$$

The sparsity of the review data means that often there are only very few reviews that have significant topic overlap with user i, so that the rank may become zero. In this case, weighting the review scores by topic profiles makes no sense. We characterize the amount of available data, and thus the confidence in the ranking, by an additional parameter $\gamma(i, A)$ that we define as the number of reviews $r_{j,A}$ with $|Z_{i,r_{j,A}}| \geq 3$.

In TPCF, we use a threshold parameter γ_0 to define the minimum confidence required for a personalized recommendation. When computing a recommendation for user i, we use $TPScore_{i,A}$ whenever $\gamma(i, A) \geq \gamma_0$ and the non-personalized average otherwise:

$$s\bar{r}_A = \frac{\sum_{j<|R_A|} sr_{j,A}}{|R_A|} \tag{9}$$

We now show the results of our experiments using this method.

4 Experimental Results

We created four experiments. The first quantifies the connection between review texts and numeric ratings. A second employs a traditional evaluation method - the mean average error, to evaluate the quality of our personalized recommendations. A third provides evidence that the proposed topic profile is coherent with the faceted opinions within the reviews, using the direct profile evaluation method. The fourth shows that using topic profiles in collaborative filtering methods increases the number of successful predictions for users showing strong preferences.

4.1 Agreement between Rating and Review Text

We first determine whether, for each individual, their rating derived preferences are correlated with the opinions in the text they write. The used dataset consisted of $|R| = 68049$ TripAdvisor[1] reviews, written by $|U| = 59067$ users for $|H| = 216$ hotels from the same location - Las Vegas, obtained with permission from TripAdvisor. From the reviews, we isolated the users U^2 with at least two hotels rated, $i \in U^2 \subset U$. For each user i, we consider every hotel pair A and B that she has rated. We use the total subjectivity of user i's review, as defined in Equation 4, as an evaluation of the sentiment expressed in the text, and consider the correlation of the difference in rating with the difference in sentiment:

$$sign(sr_{i,A} - sr_{i,B}) \cdot sign(subj(r_{i,A}) - subj(r_{i,B})) > 0 \tag{10}$$

We use the notation $cor_{subj}(\delta)$ for the percentage where we observe positive correlation (Equation 10) for review pairs whose ratings differ by a minimum amount, $|sr_{i,A} - sr_{i,B}| \geq \delta$, $\delta \in 1..4$.

The black bars in Figure 2 show the results. We notice a significant average overlap of 84.32% between the textual opinions and the rating derived preference. Moreover, the higher the rating difference is, the more supporting evidence is found in the accompanying text. In nearly 90% of cases where someone strongly prefers an item over another, the text they write will unambiguously signal that preference.

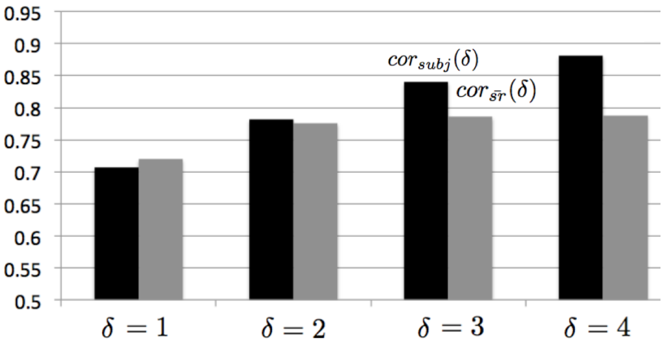

Figure 2: Rating preference correlation with textual opinions $cor_{subj}(\delta)$ (black) and other people's ratings $cor_{\bar{sr}}(\delta)$ (grey)

This correlation shows that on average ratings are consistent with the accompanying review text as extracted using our opinion extraction method.

Viewed in context, the correlation is even more meaningful. We compare it to the correlation between the rating based preference of one user and the average rating scores for those products. We compute the average of all reviews for product k, $s\bar{r}_k$. We determine for which of the users $i \in U^2$:

$$sign(sr_{i,A} - sr_{i,B}) \cdot sign(s\bar{r}_A - s\bar{r}_B) > 0 \tag{11}$$

Let $cor_{\bar{sr}}(\delta)$ be the percentage of cases for which the inequality in Equation 11 is satisfied, with $|sr_{i,A} - sr_{i,B}| \geq \delta$, for different values of δ. The grey bars from Figure 2 show the results. We find that for weak preferences (small rating differences) the predictive power of the average rating is similar to that of the user's own opinions. However for strong preferences, with $\delta \in \{3, 4\}$, the opinion based method outperforms. We have thus shown that the preferences people have, as measured through their own ratings, are consistent with what they write. This correlation is higher than that found between their preference and the ratings of others.

4.2 Relevance of Topic Profiles

We can directly evaluate the match between a topic profile and what the users actually give opinions on. The assumption is that people leave opinions about the topics they care about, more than about topics that are irrelevant to their interests.

We consider a product A, for which the target user i, with a topic profile Z_i, has written a review $r_{i,A}$. We denote with $R_{i,A}$ the set of other reviews $r_{j,A}$ for product A having $sr_{j,A} = sr_{i,A}$. If the star rating is high, we expect the subjectivity of $r_{i,A}$ with respect to Z_i to be higher than for other reviews $r_{j,A} \in R_{i,A}$. If, on the other hand, $sr_{i,A}$ is low, we expect the opposite - a relevant opinion more negative than average in the own text than in the others.

To define high and low star ratings, we compare to the neutral review $\bar{sr}$. (e.g. in the case of hotel reviews, with ratings from 1 to 5, the neutral rating is 3). We discard the reviews that have a neutral rating. Let U_A be a set of users that rated product A. We construct the subset of users $CU_A \subset U_A$ for which the following inequality holds, which we name consistent users, as their text based opinions are consistent with

[1] www.tripadvisor.com

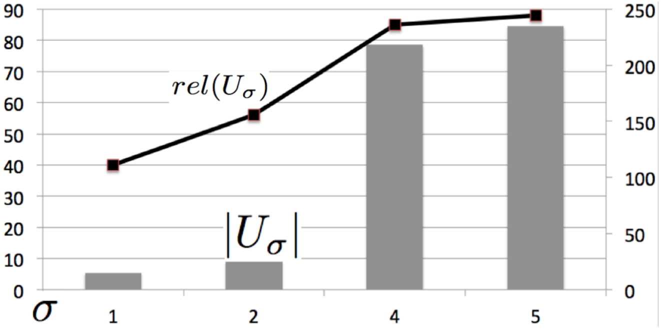

Figure 3: Profile relevance with respect to the numeric rating. Left axis: the relevance in percentage points. Right axis - the number of considered reviewers

their ratings:

$$(subj(Z_i, r_{i,A}) - \frac{\sum_{j\in R_{i,A}} subj(Z_i, r_{j,A})}{|R_{i,A}|}) \cdot sign(sr_{i,A} - \bar{sr}) > 0 \tag{12}$$

We define the relevance of a profile building method over U_A as the ratio of consistent users. The definition can afterwards be extended over various user and product sets:

$$rel(U) = \frac{\sum_{U_A \subset U} |CU_A|}{\sum_{U_A \subset U} |U_A|} \tag{13}$$

For the users $i \in U^2$, defined in the first experiment, we construct the topic profiles using the method described in Section 3.2. We test whether their high rated reviews $sr_i \in \{4, 5\}$ contain more positive opinions given the writer's important topics than average reviews. For low rated ones, $sr_i \in \{1, 2\}$ we test the opposite assumption.

In Figure 3 the bars, with numeric values on the secondary axis, show the number of reviewers $|U_\sigma|$, distributed by their rating $\sigma \in \{1, 2, 4, 5\}$, except the neutral rating. The line, corresponding to the main vertical axis, shows the consistency ratio $rel(U_\sigma)$ for each rating value.

There are two major results. The first one is that $rel(U_{\sigma\in\{4,5\}}) > 87\%$, while $rel(U_{\sigma\in\{1,2\}}) = 50\%$. This shows that, in positive reviews, people are unambiguous and the positive textual opinions abound. Negative reviews, on the other hand, display both positive and negative opinions on the important topics. A second result is that, overall $rel(U_{\sigma\in\{1..5\}}) = 83.56\%$. This shows that, for a large majority of users, the opinions they give regarding their important topics is correlated with the rating they leave.

4.3 Mean Average Error Reduction

In the following experiment, we quantify the TPCF improvements over standard rating aggregation, using a traditional evaluation - the mean average error (MAE). The dataset contains all the 68049 hotel reviews.

For each target user i, let $r_{i,A}$ be her review of hotel A, with a numeric rating sr_{i_A}. We extract user i's preferences Z_i from the review we currently analyze $Z_i = \{z \in Z | count(z, r_{i,A}) > 0\}$. To reduce randomness, we select those reviews with opinions on at least three topics $|Z_i| \geq 3$. We then compute $\gamma(i, A)$, which counts how many other reviews for hotel A, $r_{j,A}$ have $|Z_{i,r_{j,A}}| \geq 3$. If $\gamma(i, A) > \gamma_0$, then we can compute a personalized recommendation for user i. Let $R_\gamma \subset R$ be the set of all reviews, for all hotels, $r_{i,A}$ that meet this criterion and R_A all the reviews for hotel A. The bars in Figure 4, with values on the secondary axis, show the size of R_γ, for different values of γ_0.

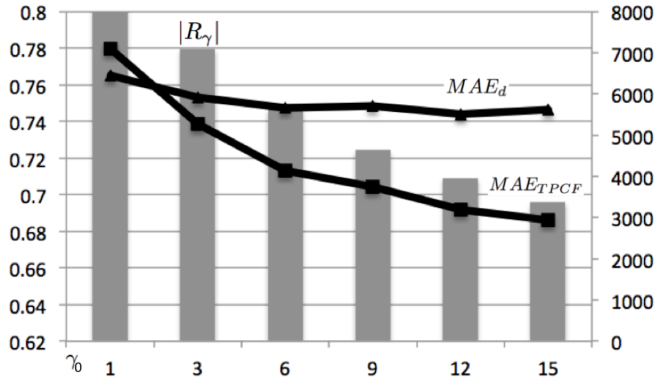

Figure 4: MAE Improvement. Left axis - MAE values. Right axis - the number of considered reviews

For each $r_{i,A} \in R_\gamma$, we define the personalized rating prediction as the score presented in Equation 8. The personalized recommendation MAE becomes

$$MAE_{TPCF} = \frac{\sum_{A\in H} \sum_{r_{i,A}\in R_A \cap R_\gamma} |sr_{i,A} - TPScore_{i,A}|}{\sum_{A\in H} |R_A \cap R_\gamma|} \tag{14}$$

The default, non personalized prediction, relies on the average star rating of hotel A, $s\bar{r}_A$:

$$MAE_d = \frac{\sum_{A\in H} \sum_{r_{i,A}\in R_A \cap R_\gamma} |sr_{i,A} - s\bar{r}_A|}{\sum_{A\in H} |R_A \cap R_\gamma|} \tag{15}$$

The lines in Figure 4 show the two MAE, with respect to γ_0. We observe that for $\gamma \geq 3$ the personalized recommender outperforms. We reduce the MAE_d error by 8% for $\gamma_0 = 15$. This result is meaningful - the more trustworthy reviews you have for a user's interests, the more relevant our recommendation is. The method only applies to roughly one tenth of all the reviews in our corpus. This current drawback is however eclipsed by the fact that currently less than 1% of people actually write reviews for the hotels they visit. Applying the TPCF method incentivizes them to write more and thus increases its own applicability. Moreover, if the method were applied on all the Tripadvisor data, there would probably be a much higher percentage of users with a high γ, leading to significantly better results.

4.4 Ranking Evaluation

The actual quality of recommendations perceived by the user is related not to the prediction of the rating, but to the relative *ranking* of different items. We should thus measure how well the ranking computed by a recommender agrees with the user's own ranking as given by her own review scores. Common measures for comparing rankings, such as Kendall's tau rank correlation coefficient [Kendall, 1938], measure this quality through the fraction of pairs with the same or different order in both rankings. In our case, for each user that has ranked several items, we can thus evaluate the quality of the ranking computed by a recommender by the fraction of times that the order of a pair of items agrees or disagrees with that given by the user herself. Instead of the Kendall correlation coefficient, we just focus on the fraction of pairs that are predicted incorrectly, averaged over all users, which we call *preference inversions*.

For two products A and B let $sr_{i,A} \neq sr_{i,B}$ be the two star ratings corresponding to the reviews left by user i. To make a recommendation, most systems compute a utility value for each alternative, and recommend the products with the highest utility values. We denote with $util_{i,A}$ and $util_{i,B}$ the utility values of products A and B, computed for user i. The ranking of the pair A and B is successful if:

$$sign(sr_{i,A} - sr_{i,B}) \cdot sign(util_{i,A} - util_{i,B}) > 0 \qquad (16)$$

We investigate the predictions made for all the users in the dataset which rated at least two products, $i \in U^2 \subset U$. We consider all pairs of products A and B that user i rated and where $sr_{i,A} \neq sr_{i,B}$. We restricted the analysis to cases where both reviews contained relevant opinions and shared at least one topic.

We impose a minimum value for the rating difference, which symbolizes the minimum preference strength $|sr_{i,A} - sr_{i,B}| > \delta$, with δ defined a priori. In the case of hotels rated from 1 to 5, $\delta = 4$ means that one hotel is rated 1 and the other 5, symbolizing a very strong preference.

Let U^2_δ be the set of users that have written pairs of reviews that meet the criteria above for a given δ value. For each user $i \in U^2_\delta$, let $sp(i)$ be the number of successfully ranked pairs, $np(i)$ the number of non-successfully ranked pairs. We define a recommender system's relevance as its aggregate performance over the analyzed users:

$$\tau(\delta) = \frac{\sum_{i \in U^2_\delta} sp(i) - np(i)}{\sum_{i \in U^2_\delta} sp(i) + np(i)} \qquad (17)$$

For $\delta = 1$, the recommender relevance is equal to the Kendall tau rank correlation coefficient [Kendall, 1938].

Let $\tau_{TPCF}(\delta)$ be the relevance of the TPCF method. We compute $TPScore_{i,A}$ and $TPScore_{i,B}$ as in Equation 8 and then define successful TPCF predictions as cases where

$$sign(sr_{i,A} - sr_{i,B}) \cdot sign(TPScore_{i,A} - TPScore_{i,B}) > 0 \qquad (18)$$

which is a particular case of Equation 16.

We compare the TPCF results with those of the unpersonalized, default recommender. The default recommender makes a successful prediction if the inequality in Equation 11 holds. We use $\tau_{\bar{sr}}(\delta)$ to show the relevance of the default recommender system.

Figure 5 presents the evaluation results. We notice that the predictions made by the TPCF method underperform for low strength preference pairs: $\tau_{TPCF}(\delta) < \tau_{\bar{sr}}(\delta), \delta \leq 2$. However, if the preference is strong ($\delta \geq 3$), the TPCF relevance is significantly higher than the benchmark: $\tau_{TPCF}(\delta) = 3 \cdot \tau_{\bar{sr}}(\delta), \delta = 4$. The proposed evaluation is helpful in comparing personalized and non personalized recommenders. Note that due to the lack of data, the experiment was carried out with $\gamma_0 = 4$, a threshold that on average shows only very small improvement in MAE. We believe that in a denser dataset, where γ_0 can be set to a higher value, we would see a more significant difference in ranking quality even for lower rating differences.

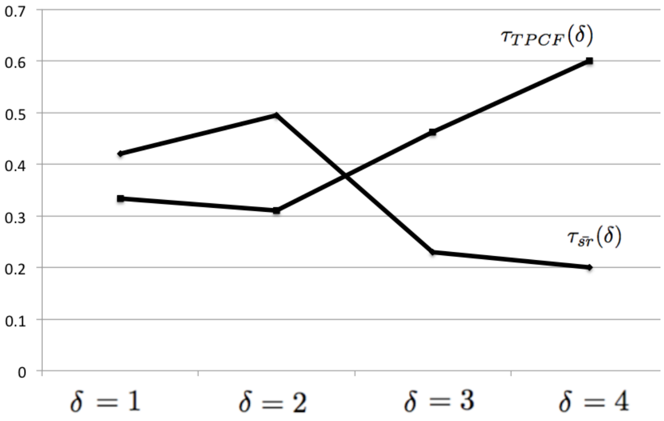

Figure 5: $\tau_{TPCF}(\delta)$ and $\tau_{\bar{sr}}(\delta)$ for different δ values

5 Conclusion and Future Work

In this paper, we considered the problem of personalizing rankings given by review sites. To address the problem posed by the extreme sparsity of this data, we proposed a new method called topic profile collaborative filtering that uses the topics present in reviews written by a user to filter the reviews used for a recommendation.

We verified the underlying intuition that review texts correlate with ratings and that topic profiles point to the relevant opinions, and then showed that the technique indeed produces a lower mean average error in rating prediction and a better predictions of users' preference rankings. We believe that Kendall's tau correlation coefficient applied to the ranking of items where a user's evaluation is known is a more meaningful measure of accuracy than the common mean average error of ratings, since it more closely reflects the accuracy of what is shown to a user.

A strong limitation of our work is that we did not have access to the complete data available on the review site, and thus did not always find enough similar reviews to make meaningful personalized predictions. Thus, we had to condition our positive results on the cases where enough data was indeed available. We hope to be able to evaluate this technique on a denser, more complete data such where we believe it will produce much more significant results.

References

[Balabanović and Shoham, 1997] Marko Balabanović and Yoav Shoham. Fab: content-based, collaborative recommendation. *Commun. ACM*, 40(3):66–72, March 1997.

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[de Marneffe *et al.*, 2006] Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure trees. In *LREC*, 2006.

[Faridani, 2011] Siamak Faridani. Using canonical correlation analysis for generalized sentiment analysis, product recommendation and search. In *Proceedings of the fifth ACM conference on Recommender systems*, RecSys '11, pages 355–358, New York, NY, USA, 2011. ACM.

[Hariri *et al.*, 2011] N. Hariri, B. Mobasher, R. Burke, and Y. Zheng. Context-aware recommendation based on review mining. In *Proceedings of the 9th Workshop on Intelligent Techniques for Web Personalization and Recommender Systems (ITWP 2011)*, page 30, 2011.

[Homoceanu *et al.*, 2011] Silviu Homoceanu, M. Loster, Christoph Lofi, and Wolf-Tilo Balke. Will i like it? ? providing product overviews based on opinion excerpts. In *Proc. of IEEE Conference on Commerce and Enterprise Computing (CEC)*, 2011.

[Jakob *et al.*, 2009] Niklas Jakob, Stefan Hagen Weber, Mark Christoph Müller, and Iryna Gurevych. Beyond the stars: exploiting free-text user reviews to improve the accuracy of movie recommendations. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, TSA '09, pages 57–64, New York, NY, USA, 2009. ACM.

[Kazienko and Kolodziejski, 2006] Przemyslaw Kazienko and Pawel Kolodziejski. Ko?odziejski p.: Personalized integration of recommendation methods for e-commerce. int. *Journal of Computer Science and Applications*, 3, 2006.

[Kendall, 1938] M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1/2):81–93, June 1938.

[Levi *et al.*, 2012] Asher Levi, Osnat Mokryn, Christophe Diot, and Nina Taft. Finding a needle in a haystack of reviews: cold start context-based hotel recommender system. In *Proceedings of the sixth ACM conference on Recommender systems*, RecSys '12, pages 115–122, New York, NY, USA, 2012. ACM.

[Middleton *et al.*, 2004] Stuart E. Middleton, Nigel R. Shadbolt, and David C. De Roure. Ontological user profiling in recommender systems. *ACM Trans. Inf. Syst.*, 22(1):54–88, January 2004.

[Miller, 1995] George A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[O'Mahony and Smyth, 2010] M. P. O'Mahony and B. Smyth. A classification-based review recommender. *Know.-Based Syst.*, 23(4):323–329, May 2010.

[Pazzani and Billsus, 1997] Michael Pazzani and Daniel Billsus. Learning and revising user profiles: The identification ofinteresting web sites. *Mach. Learn.*, 27(3):313–331, June 1997.

[Pitsilis and Knapskog, 2012] Georgios Pitsilis and Svein J. Knapskog. Social trust as a solution to address sparsity-inherent problems of recommender systems. *CoRR*, abs/1208.1004, 2012.

[Poirier *et al.*, 2010] Damien Poirier, Isabelle Tellier, Françoise Fessant, and Julien Schluth. Towards text-based recommendations. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, pages 136–137, Paris, France, France, 2010. Le Centre de Hautes Etudes Internationales D'Informatique Documentaire.

[Sarwar *et al.*, 2001] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 285–295, New York, NY, USA, 2001. ACM.

[Schein *et al.*, 2002] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, pages 253–260, New York, NY, USA, 2002. ACM.

[Shani and Gunawardana, 2009] Guy Shani and Asela Gunawardana. Evaluating recommender systems - microsoft research. Technical report, November 2009.

[Snyder and Barzilay, 2007] Benjamin Snyder and Regina Barzilay. Multiple aspect ranking using the good grief algorithm. In *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL*, pages 300–307, 2007.

[Sugiyama *et al.*, 2004] Kazunari Sugiyama, Kenji Hatano, and Masatoshi Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 675–684, New York, NY, USA, 2004. ACM.

[Wilson *et al.*, 2005] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Opinionfinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, pages 34–35, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[Zhang *et al.*, 2010] Zi-Ke Zhang, Chuang Liu, Yi-Cheng Zhang, and Tao Zhou. Solving the cold-start problem in recommender systems with social tags. *EPL (Europhysics Letters)*, 92(2):28002, 2010.