The Thirty-First International Florida
Artificial Intelligence Research Society Conference (FLAIRS-31)

A Multi-Domain Analysis of Explanation-Based Recommendation Using User-Generated Reviews

Khalil Muhammad, Aonghus Lawlor, Barry Smyth
Insight Centre for Data Analytics
University College Dublin, Belfield, Dublin 4, Ireland
{khalil.muhammad, aonghus.lawlor, barry.smyth}@insight-centre.org

Abstract

This paper extends recent work on the use of explanations in recommender systems. In particular, we show how explanations can be used to rank as well as justify recommendations, then we compare the results to more conventional recommendation approaches, in three large-scale application domains.

Introduction

There is an increasing need for AI systems to justify and explain their decisions to end-users, as institutions such as the EU roll out regulations that contain a *right to explanation* for its citizens. While such regulations will introduce significant challenges for an increasingly data-driven algorithmic world, they also motivate interesting questions and present new opportunities for AI research. In this paper, we consider the importance of explanations in recommender systems.

Researchers have long understood the importance of explanations to justify these recommendations. Early work often focused on different styles of explanation interfaces – how explanations should be presented and how users perceive them (Herlocker, Konstan, and Riedl 2000) – intending to improving transparency, persuasiveness, and trust. Recently, (Musto et al. 2016) generated natural language explanations using information from the Linked Open Data, and (Chang, Harper, and Terveen 2016) employed a review-based approach to explain movie recommendations.

We will argue in favour of deeper explanations, which more authentically convey the true reasons behind recommendations. More relevant here is research by (Muhammad et al. 2015; Muhammad, Lawlor, and Smyth 2016) which has argued for a more intimate connection between recommendation and explanation, by using explanations to rank recommendations. Specifically, we extend recent work in this area by comparing a number of different approaches to such *explanation-based recommendation* and evaluating their effectiveness and comparing their relative performance in three different recommendation domains. Our findings show explanation-based recommendation to be more effective than conventional content-based and collaborative filtering approaches, but the benefits observed depend on certain domain characteristics and review properties.

Explanation-Based Ranking

Earlier work by (Muhammad et al. 2015) proposed a way to estimate the *strength* of an explanation and speculated that this score could be used to rank items for recommendation. Thus, items that are associated with stronger or more compelling explanations should be ranked ahead of items with weaker, less compelling explanations. We extend this work with a *weighted* form of explanation strength, such that each *pro* and *con* is associated with a weight (w_f) to indicate the relative importance of the feature as shown in Eq 1. By weighting features in this way we can adjust the strength score of an explanation based on whether or not its *pros* or *cons* are particularly important — to the user or the item.

$$Strength(u_T, i, I') = \sum_{f \in Pros(u_T, i, I')} w_f * better(f, i, I') - \sum_{f \in Cons(u_T, i, I')} w_f * worse(f, i, I') \quad (1)$$

Here, $better(f, i, I')$ measures how a target item i is *better* than alternative recommendations I' based on feature f; $worse(f, i, I')$ measures how i is *worse* than alternative recommendations I' based on feature f. Thus, recommendations associated with explanations that are predominantly positive — more, important *pros* with higher *better* scores and fewer, less important *cons* with lower *worse* scores — will have a high *strength* score. Such explanations should offer the user a better choice of recommendation, with fewer compromises with respect to the features that matter to them. By contrast, recommendations that are associated with a lower or even negative *strength* score will usually involve far more compromises from the user in terms of the features that matter to them.

We consider three variations of this scoring metric as the basis for explanation-based ranking (EBR). In the first we use *uniform* weights ($w_f = 1$), which is equivalent to that proposed by (Muhammad et al. 2015). In this variation, each feature is just as important as every other feature and the strength score depends purely on whether the sentiment of that feature in the current item is better or worse than that feature‘s sentiment in the alternative recommendations.

Clearly, the features should not all be treated equally. Some features occur in a great many reviews of an item —

they have a high item *importance* score — indicating that these features are important in the context of this item. For example, a particular hotel may pride itself on its *leisure centre*, and so we might expect to see this feature frequently mentioned in reviews of the hotel. Thus, in our second variation of the strength metric we weight features by their item *importance* score. Equation 2 shows how to compute the importance score for entity e (a user or item), where $R(e)$ is the set of all features mined from the reviews of e.

$$imp(f_j, e) = \frac{count(f_j, i)}{\sum_{f' \in R(e)} count(f', i)} \tag{2}$$

Another approach is to focus on features that are important to the target user. The intuition here is that if user u frequently mentions *customer service* in their restaurant reviews, then it makes sense to give more weight to this feature in the strength score of this explanation. We can do this by using the user *importance* scores to weight features (see 2).

Evaluation

We evaluate different forms of EBR (*uniform*, *item* and *user* weights) on real-world datasets, and in comparison to content-based and collaborative filtering recommendation methods. We use real-world review datasets from BeerAdvocate, Yelp, and TripAdvisor. Each dataset occupies a different position in the review-space and is essentially made up of a set of review-tuples, (r, u, i), with a review r, written by user u, for item i; as summarised in Table 1. On average, a typical BeerAdvocate (BA) item is associated with 4 features, and a user profile contains just 5 features. TripAdvisor profiles contain 3 features, and a typical item is associated with 11 features, on average. In Yelp (YP), a typical item just over 3 features and a typical profile about 4 features.

Data	Reviews	Items	Users	Reviews per Item	Reviews per User
BA	131,418	17,856	5,710	8 ± 21.6	23 ± 83.6
TA	43,528	1,982	10,000	22 ± 24.1	4 ± 1.93
YP	23,109	8,048	10,000	3 ± 3.9	2.3 ± 4.1

Table 1: Datasets used in evaluating EBR.

Methodology

For each review-tuple, (r, u, i), our datasets also contain the 10 best recommendations $(i_0, ..., i_9)$ that are typically made alongside the item i (by TripAdvisor, Yelp, or BeerAdvocate) per user. Thus, we can consider each review-tuple to define a recommendation session in which some target user u is looking for some item i — one that they consumed in the past since they wrote a review about it — and is presented with a list of suggestions, $i_0, ..., i_9$. Each of these suggestions has an overall rating score, and the items are ranked using this score. We will treat each of these ratings-based rankings as the *ground-truth* against which to judge the quality of the rankings produced by our test algorithms. For each review-tuple, we will generate a set of alternative rankings for the 10 suggested items, and we will compare these rankings to the ground-truth in various ways.

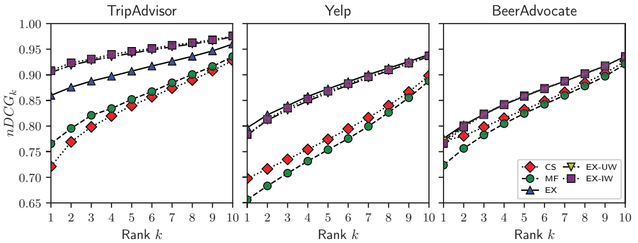

Figure 1: Average $nDCG@k$ per session against rank position for both explanation-based similarity-based, matrix factorisation based ranking methods

Ranking Algorithms When it comes to testing alternative rankings, we will compare five different algorithms: three EBR variations (based on *uniform*, *user*, or *item* weights), and two benchmark algorithms using more conventional recommendation techniques. We will refer to *user* weights as UW (or $EX - UW$), *item* weights as IW (or $EX - IW$), and *uniform* weights as NW (or EX).

The first of the benchmark algorithms reflects a common similarity-based approach to ranking in which a set of items are ranked based on how similar they are to a user's profile (Singhal 2001). We refer to this technique as CS because we rely on a cosine similarity metric, which compares the features of an item to the features of a user profile to determine a similarity score as the basis for ranking.

The second benchmark algorithm adopts a state-of-the-art collaborative filtering Matrix Factorisation (MF) technique to predict the ratings for the items for a target user and then rank these items based on their predicted ratings. In each of the above, whether explanation-based, content-based, or collaborative, we generate an alternative ranking of the suggested items $i_0, ..., i_9$ for a given target user (u).

On Ranking Quality We will compare the each of the rankings produced by the five test algorithms to the ground-truth. Note that we are using this as a ground-truth not so much because it represents an ideal ranking per se, but rather because it represents a useful ranking for these systems currently in use. To compare one ranking to the ground-truth, we use a normalised discounted cumulative gain metric to compare the average rating of the items up to each rank positions. Thus, *nDCG@k* refers to the $nDCG$ value for items up to and including position k; an $nDCG$ of 1 means that the ranking is identical (by ratings) to the ground-truth.

The results, for each domain, are presented in Fig 1. In each graph we show the $nDCG_k$ line (for $k = 1...10$) for each of the 3 explanation-based techniques (*uniform, user, item*) as well as *CS* and *MF*. For instance, we see the results for the TripAdvisor dataset and it is clear that, on average, all of the explanation-based techniques present with higher $nDCG$ scores than CS and MF across all values of k. In other words, the explanation-based method of ranking produces results that are much closer to the ground-truth based on overall item ratings than the CS and MF rankings. This pattern is evident in the Yelp and BeerAdvocate datasets too.

However, we can also see that, except in TripAdvisor, us-

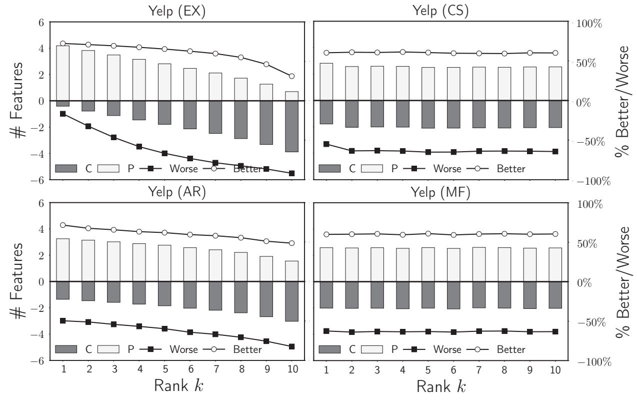

Figure 2: Average number of *pros*/*cons*, *better*/*worse* for Yelp restaurants ranked by CS and $Uniform$ approaches.

ing user weights (UW) or item weights (IW) during the calculation of explanation strength does not lead to rankings that are materially different from those obtained using uniform unweighted EBR (i.e. EX). This may be linked to the fact that the Yelp and BeerAdvocate datasets have relatively few features per item, compared to TripAdvisor. A detailed discussion of this is beyond the scope of this paper.

In summary then, these results show how EBR methods are capable of producing rankings that are closer to the ground-truth rankings, compared with CS or MF methods. But, as mentioned earlier it is not strictly true to claim that the rating-based (AR) ranking provides a true ground-truth. Ranking items by their overall ratings is a reasonable ground-truth if users are likely to be swayed by higher ratings; the potential exists to produce better rankings, especially given that an overall rating for an item may not tell the full story. In the following sections, we explore this by using the explanations of items to evaluate different rankings.

Explanations by Rank We can generate an explanation for each item in any recommendation session. These explanations serve as a summary of how positive or negative a given item is likely to be for the target user: more *pros* and fewer *cons* makes for a stronger explanation and a more positive item recommendation involving fewer compromises.

It is useful, therefore, to compare the explanations that are associated with the different rank positions in the rankings produced by the various techniques. For each rank position, we can calculate the average number of *pros* and *cons*, and the average *better*/*worse* scores, for the corresponding explanations. Ideally, we should expect to see items with more positive explanations (more *pros* and fewer *cons*) to be ranked higher than items with more negative explanations (fewer *pros* and more *cons*). The results for the Yelp domain are shown in Fig 2 for EX, AR, CS, and MF. It appears that there is a material difference between the rankings produced by EX and AR compared with CS and MF. In the case of EX, we can see how more highly ranked items are associated with explanations that have more *pros* (with higher *better* scores) and fewer *cons* (with lower *worse* scores); suggesting more positively reviewed items with fewer compromises. As we move down the ranking the number of *pros* (and their *better* scores) decreases, while the number of *cons* (and their *worse* scores) increase. Items in the lower half of the ranking involve far more compromises than items in the upper half of the ranking. Similar patterns were observed in BeerAdvocate and TripAdvisor domains but were left out due to space constraint.

Ideally, the EBR approach should produce this effect since the items are ranked by explanation strength, which is a function of the *pros*/*cons* and *better*/*worse* scores. However, it is not guaranteed for the AR approach to produce a similar effect, and yet it does. This speaks to the usefulness of the AR ranking, at least when it comes to showing more positive items at the top of the ranking and more negative items at the bottom. We see a very different effect for CS and MF, however; neither of them appears to be able to sort the more positive items from the less positive items. In each case, regardless of domain, there is little to differentiate between items at the top and bottom of the recommendation lists, at least regarding *pros* and *cons*.

The point here is that both explanation-based and AR rankings produce recommendation lists which prioritise more positive items, while CS and MF rankings do not. The AR approach, used by TripAdvisor, Yelp, and BeerAdvocate does this based on overall item-level ratings — which are not necessarily correlated to the sentiment expressed in reviews for various reasons — provided separately by users, whereas our explanation-based technique does this based on the sentiment of individual features from reviews.

Explanation Strength In Fig 2, the difference between AR and EX rankings, and those produced by CS and MF is clear to see, but is there a significant difference between EX and AR? The strength of an explanation is designed to measure how compelling an explanation is, and as such it can provide a useful way to evaluate the explanations associated with items at the various rank positions using different ranking strategies. Fig 3 shows the strength scores (Eq. 1) for the explanations associated with recommendations at each rank position based on the rankings produced by AR, CS, MF and EX. Each row of the grid corresponds to one of the three weighting approaches, and each column refers to a specific domain. The results clearly show how the explanation-based approaches EX, and to a lesser extent AR, produce rankings that are sensitive to the overall strength of the corresponding items. This is not surprising in the case of the explanation-based approaches, after all, they are ranked by explanation strength to begin with, but the extent of the difference between these rankings and AR, not to mention CS and MF, is striking. For each dataset, regardless of weighting model used in EBR, the EX exhibit a strong ranking signal, which sees items with the most positive overall sentiment being ranked ahead of those with a less positive sentiment. The same is broadly true with the AR rankings, although the difference in strength across the rank positions is less obvious. However, this ranking signal is not evident among the CS and MF rankings and, as mentioned in the previous section, there is little to distinguish the top-ranked items from the bottom-ranked items in terms

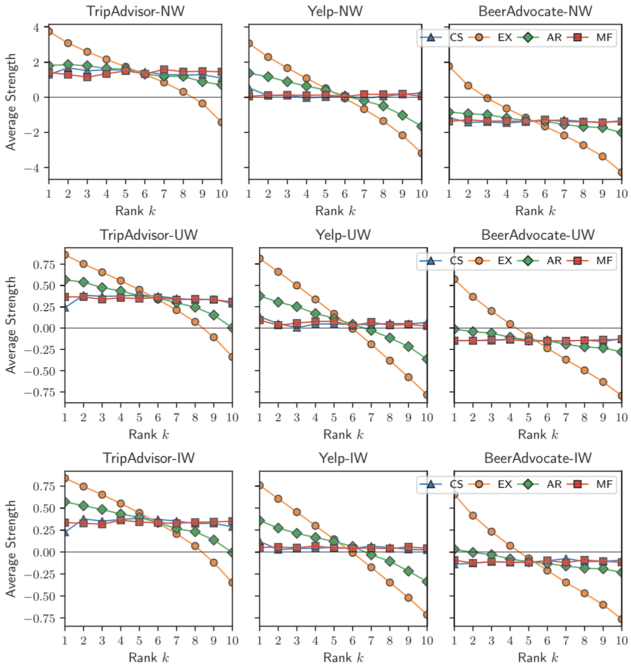

Figure 3: Strength per rank position of different variations.

of their explanation strength. Importantly, we can now see a clear difference between the explanation-based rankings EX and the AR ranking, which is serving as a ground-truth. For every rank position (from position 1 to 5), EBR methods can identify items whose explanations have a significantly higher strength score than those identified by AR. For lower rank positions, AR tends to have higher rank scores, suggesting that it is incorrectly ranking some positive items too far down the ranking.

This analysis demonstrates how explanation-based methods are capable of producing better rankings, assuming the mined features and sentiments are an accurate reflection of review sentiment and user opinion. Because the explanation-based methods operate at the level of individual features, rather than the summary ratings score used by AR, they are able to make more fine-grained distinctions between items.

Discussion and Limitation

EBR, unlike CS, prioritises items that are predominantly positive with few negatives. Moreover, the results from EBR tend to be closer to the ideal ranking compared to CS, suggesting a better overall recommendation session. The benefits of EBR appear to be more significant in TripAdvisor and Yelp, but less so in BeerAdvocate. Including user and item weights significantly improves the recommendations in TripAdvisor, there is a small improvement in the Yelp domain and hardly any in with BeerAdvocate. The reasons for these differences hinge on the characteristic nature of reviews (i.e., their features and sentiment), across the domains. One factor that is likely to have an impact is that in TripAdvisor, we use a greater number of features (19) in the item and user profiles compared to BeerAdvocate (6) and Yelp (8). TripAdvisor users and items are distinguished by different mixes of features with different weightings. In contrast, BeerAdvocate and Yelp reviews draw on a more limited set of features, which makes for a more homogeneous user and item descriptions. Answering the question why the user or item weights hardly improve the ranking in Yelp or BeerAdvocate will require additional work and is a future research priority.

One limitation of our evaluation is that there no live user study. A live user study on this scale will be difficult and expensive to conduct. Another limitation is the choice of baselines used in the evaluation. We use two baselines in the form of similarity-based and collaborative filtering approaches in line with current best practice. We use the review ratings as a ground-truth for our evaluation because this is what the major review sites use today. Our $nDCG$ results show that our explanation approach is closer to this ground-truth than similarity or matrix factorisation techniques. A further validation of our approach is that we can produce rankings which are different from the ratings-based approaches, making better use of more granular opinions than simple ratings.

Conclusions

We extend recent work in this area of recommender systems which used explanations to drive the recommendation process. We described and evaluated a number of different explanation-based ranking techniques and compared their performance to suitable content-based and collaborative filtering baselines in three different domains. Our study of different approaches to weighting the relative *importance* of explanation features has been less conclusive but highlights differences across the review domains. Future work will strive to better understand the domain characteristics that are likely to determine the extent that explanation-based techniques can work.

Acknowledgements. This research is supported by Science Foundation Ireland under grant number SFI/12/RC/2289.

References

Chang, S.; Harper, F. M.; and Terveen, L. 2016. Crowd-Based Personalized Natural Language Explanations for Recommendations. In *RecSys '16*, 175–182. ACM.

Herlocker, J. L.; Konstan, J. A.; and Riedl, J. 2000. Explaining Collaborative Filtering Recommendations. In *CSCW '00*, 241–250.

Muhammad, K.; Lawlor, A.; Rafter, R.; and Smyth, B. 2015. Great Explanations: Opinionated Explanations for Recommendations. In *ICCBR '15*. 244–258.

Muhammad, K.; Lawlor, A.; and Smyth, B. 2016. A Live-User Study of Opinionated Explanations for Recommender Systems. In *IUI '16*, 256–260. Sonoma, USA: ACM.

Musto, C.; Narducci, F.; Lops, P.; De Gemmis, M.; and Semeraro, G. 2016. ExpLOD: A Framework for Explaining Recommendations based on the Linked Open Data Cloud. In *RecSys '16*, 151–154. ACM.

Singhal, A. 2001. Modern Information Retrieval: A Brief Overview. *IEEE Data Engineering Bulletin* 24(4):35–43.