A Live-User Study of Opinionated Explanations for Recommender Systems

Khalil Muhammad
Insight Centre for Data Analytics
University College Dublin
khalil.muhammad@insight-centre.org

Aonghus Lawlor
Insight Centre for Data Analytics
University College Dublin
aonghus.lawlor@insight-centre.org

Barry Smyth
Insight Centre for Data Analytics
University College Dublin
barry.smyth@insight-centre.org

ABSTRACT

This paper describes an approach for generating rich and compelling explanations in recommender systems, based on opinions mined from user-generated reviews. The explanations highlight the features of a recommended item that matter most to the user and also relate them to other recommendation alternatives and the user's past activities to provide a context.

ACM Classification Keywords

H.3.3 Information Search and Retrieval: Information filtering; H.5.2 User Interfaces: Evaluation/methodology

Author Keywords

Recommender Systems; Explanations; Opinion Mining

INTRODUCTION

Recommender systems attempt to learn about a user's preferences in order to make targeted suggestions about the items they may be interested in. Recently researchers have turned their attention to explaining recommendations to make it easier for users to make informed decisions, with a view to increasing conversion rates and leading to more satisfied users [1–5]. Early work explored the utility of explanations in collaborative filtering systems with [1] reviewing different models and techniques for explanation based on MovieLens data. They considered a variety of explanation interfaces leveraging different combinations of data (ratings, meta-data, neighbours, confidence scores etc.) and presentation styles (histograms, confidence intervals, text etc.) concluding that most users recognised the value of explanations.

Bilgic and Mooney [6] used keywords to justify items rather than disclosing the behaviour of similar users. They argued that the goal of an explanation should not be to "sell" the user on the item but rather to help the user to make an informed judgment. Elsewhere, keyword approaches were further developed by [2] in a content-based, collaborative hybrid capable of justifying recommendations as: *"Item A is suggested because it contains feature X and Y that are also included in items B, C, and D, which you have also liked."*; see also the work of [7] for related ideas based on user-generated tags instead of keywords. Explanations can also relate one item to others. For example, Pu and Chen [3] build explanations that emphasise the tradeoffs between items, such as *"Here are laptops that are cheaper and lighter but with a slower processor"*. In this paper we describe the results of a user study designed to evaluate the relative merits of different styles of recommendation explanations. The recommendation setting, and the explanations produced are novel in the sense that item descriptions and user profiles are mined directly from user-generated reviews.

DOI: http://dx.doi.org/10.1145/2856767.2856813

OPINION MINING FOR RECOMMENDATION

This paper builds on recent work [8–11] by the community about mining opinions from user reviews to generate user profiles and item descriptions for recommender systems. The work of [10, 11] is especially relevant and describes how shallow opinion mining techniques can be used to extract rich feature-based item descriptions (item *cases*) based on the features that users refer to in their reviews and the polarity of their opinions. An in-depth description of this approach is beyond the scope of this paper and the interested reader is referred to [10, 11] for further details. However, in the interest of what follows we will briefly summarise the type of opinion data that can be produced for the purpose of recommendation and this forms the basis for our explanations. Without loss of generality, we will do this using a reference dataset of TripAdvisor hotel reviews that is available privately for academic use only. The dataset comprises of information about 2,370 hotels, 150,961 users and 227,125 reviews from June to August 2013.

Generating Item Descriptions

Each item/hotel h_i is associated with a set of reviews $R(h_i) = \{r_1, \ldots, r_n\}$, the opinion mining process extracts a set of features, $f_1, \ldots, f_m$, from these reviews, as described in [10, 11]. Each feature, f_j is associated with an *importance* score and a *sentiment* score as per Equations 1 and 2. The importance score, $imp(f_j, h_i)$, is the relative number of times that f_j is mentioned in $R(h_i)$. The sentiment score, $s(f_j, h_i)$, is the degree to which f_j is mentioned positively or negatively in $R(h_i)$.

$$imp(f_j, h_i) = \frac{count(f_j, h_i)}{\sum_{\forall f' \in R(h_i)} count(f', h_i)} \quad (1)$$

$$s(f_j,h_i) = \frac{pos(f_j,h_i)}{pos(f_j,h_i)+neg(f_j,h_i)} \quad (2)$$

$$item(h_i) = \{(f_j, s(f_j,h_i), imp(f_j,h_i)) : f_j \in R(h_i)\} \quad (3)$$

An item/hotel case description is a representation of these features and scores as per Equation 3. Note, $pos(f_j,h_i)$ and $neg(f_j,h_i)$ denote the number of mentions of f_j labeled as positive and negative during the sentiment analysis phase. So $s(f_j,h_i)$ values close to *1* mean that positive opinions dominate whereas a score close to *0* means that negative opinions dominate.

Generating User Profiles

Similarly, we can generate a profile of a user u_T based on the reviews that they have written by extracting features and importance information from these reviews as in Equation 4.

$$user(u_T) = \{(f_j, imp(f_j,u_T)) : f_j \in R(u_T)\} \quad (4)$$

We give more meaning to the frequency with which the user reviews a particular feature, rather than the average sentiment of the user's opinions, since the frequency of mentions is a better indication of which features matter most to a user. For users that have not written any reviews, we can build their profiles by mining opinions from reviews that they have found helpful in the past, or by mining those features that tend to dominate in the hotels that they have stayed in.

FROM OPINIONS TO EXPLANATIONS

Our aim is to describe an approach for generating explanations for each item in a set of recommendations $H = \{h_i...h_k\}$ generated for some user u_T. The novelty of our approach stems from how we leverage opinion information in two ways: (1) to highlight those important features (*pros* and *cons*) of an item that likely matter to u_T; (2) to emphasise those features that distinguish the recommendation relative to other items, such as alternative recommendations or past bookings.

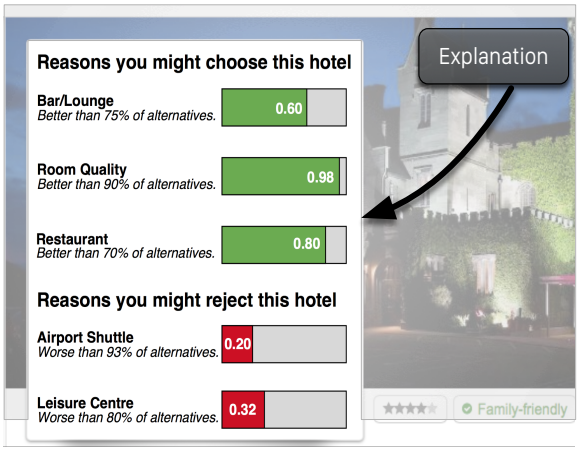

Figure 1. An example explanation showing *pros* and *cons* that matter to the target user along with sentiment indicators (horizontal bars) and information about how this item fares with respect to alternatives.

Core Explanation Components

To begin with, we present in Figure 1 an example explanation for one particular hotel. There are a number of components worth highlighting. First, the explanation is made up of a number of features that have been extracted from the reviews of this hotel and that are known to matter to the user; these are features that the user has mentioned in their own past reviews. Second, these features are divided into *pros* and *cons*, the former with positive sentiment scores $s(f_j,h_i) > 0.7$ and the latter with more negative sentiment scores $s(f_j,h_i) < 0.7$. The *pros* might be reasons to choose the hotel whereas *cons* might be reasons to avoid it. In the case of our data, there are significantly more positive sentiments than negatives, thus a sentiment threshold of 0.7 provides a reasonable split between pros and cons. Third, each feature is associated with a *sentiment bar* that shows the actual sentiment score for that feature. And finally, each feature is associated with an additional piece of explanatory text that highlights how the hotel compares to other relevant items called a *reference set* H' — such as alternative recommendations as in this example — in terms of this feature; the aim here is to provide the user with some additional explanatory context by relating the feature to other hotels that may be relevant to their decision.

	Feature	Importance	Sentiment	better/worse
PROS	Bar/Lounge *	0.25	0.71	75%
	Free Breakfast	0.22	0.79	10%
	Room Quality *	0.18	0.98	90%
	Restaurant *	0.15	0.80	80%
	Shuttle Bus	0.06	0.75	10%
CONS	Airport Shuttle *	0.21	0.20	90%
	Leisure Centre *	0.11	0.32	80%
	Swimming Pool	0.10	0.45	33%
	Room Service	0.5	0.46	20%
	:	:	:	:

Figure 2. An example of an explanation structure showing *pros* and *cons* that matter to the user along with associated importance, sentiment, and better/worse than scores.

Generating a Basic Explanation Structure

To generate an explanation like the one shown in the previous section, we start with a *basic explanation structure* that is made up of the features of the item in question (h_i) which are also present in the user's profile u_T. These features are divided into *pros* and *cons* based on their sentiment score $s(f_j,h_i)$ and ranked in order of importance $imp(f_j,u_T)$.

We also compute so-called *betterThan* (BT) and *worseThan* (WT) scores as in Equations 5 and 6 with respect to some suitable reference set H'. These scores calculate the percentage of items in the reference set for which f_j has a better sentiment score (for *pros*) or worse sentiment score (for *cons*) in h_i. Suitable reference sets include the set of alternative recommendations and the users own past bookings; in this section we assume the former.

$$BT(f'_j,h_i,H') = \frac{\sum_{h_c \in H'} 1[s(f'_j,h_i) > s(f'_j,h_c)]}{|H'|} \quad (5)$$

$$WT(f'_j,h_i,H') = \frac{\sum_{h_c \in H'} 1[s(f'_j,h_i) < s(f'_j,h_c)]}{|H'|} \quad (6)$$

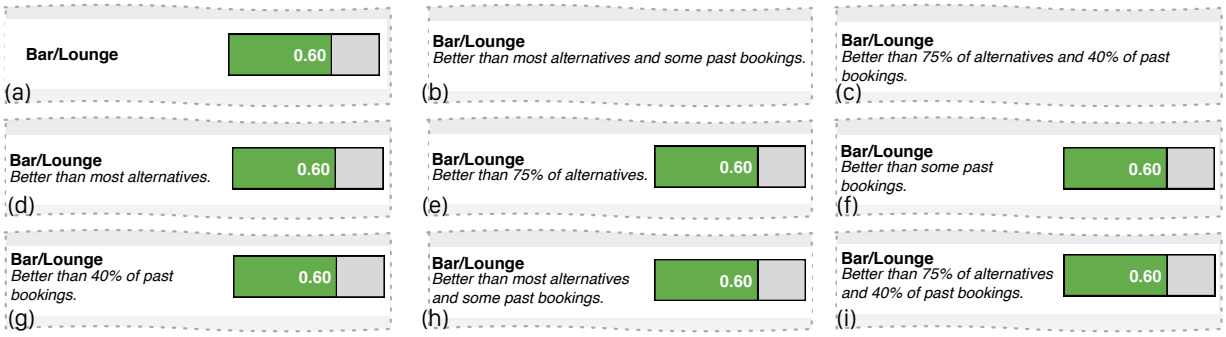

Figure 3. Fragments of explanation interfaces showing all 9 variations (a)-(i); examples of complete interfaces are shown in Figure 4.

Figure 4. Full sample explanations for the variations (a), (e) and (i).

Figure 2 shows an example basic explanation structure with set of 5 *pros* and 4 *cons*. For each we can see its importance to the user, its sentiment score from the hotel's reviews, and the corresponding better/worse scores. In this case the reference set is the alternative recommendations suggested alongside this hotel (which are not shown here). For example, we see that the *Bar/Lounge* feature, with a sentiment score of 0.6, is better than 75% of the alternative recommendations.

From Basic to Compelling Explanations

Not every feature in the previous example makes for a compelling reason to choose or reject the hotel in question. For example, the *Free Breakfast*, while positively reviewed, is only better than 10% of the alternative recommendations. If this feature is important to the user then there are better alternatives to chose from. In contrast, this hotel's *Room Quality* beats 90% of the alternatives and so may offer a strong reason to prefer this hotel. To simplify the explanations that are presented to users, and make them more compelling, we filter out features than have lower better/worse scores ($< 50\%$) so that only those features that are better/worse than a majority of alternatives remain; these features are indicated with an asterix in Figure 2. They are all features that matter to the user and they distinguish the hotel as either better or worse than a majority of comparable items, e.g. alternative recommendations.

In summary then, to produce an explanation for some recommended item h_i for user u_T given some reference set of related items H' we first identify those features of h_i that matter to u_T, based on their profile. Next we separate these features into *pros* and *cons*. And then we eliminate those *pros* that are not better than a majority of the reference set and those *cons* that are not worse than a majority of the reference set. Finally, we rank-order these remaining (compelling) features by importance to u_T for inclusion in the final explanation.

EVALUATION

We have described an approach to generating novel recommendation explanations based on opinions mined from user-generated reviews. These explanations can be generated to help justify a recommended item to the user by highlighting features that matter to them. We argue that the user can make a more informed decision when the features of the item being explained are separated into *pros* and *cons*, and then augmented with sentiment and contextual information.

Setup

We prepared a range of different explanation styles/interfaces as the initial part of a user study. All of the interfaces separated features into *pros* and *cons*, but each variation was different in terms of whether sentiment bars were used, which reference sets were chosen (alternative recommendations vs. past bookings), and whether the better/worse scores were presented as precise percentages or not. For reasons of space we cannot show the complete set of all 9 interfaces. Instead, in Figure 3 we present the relevant (single-feature) fragments from each of the 9 interfaces and the reader is also referred to Figure 4

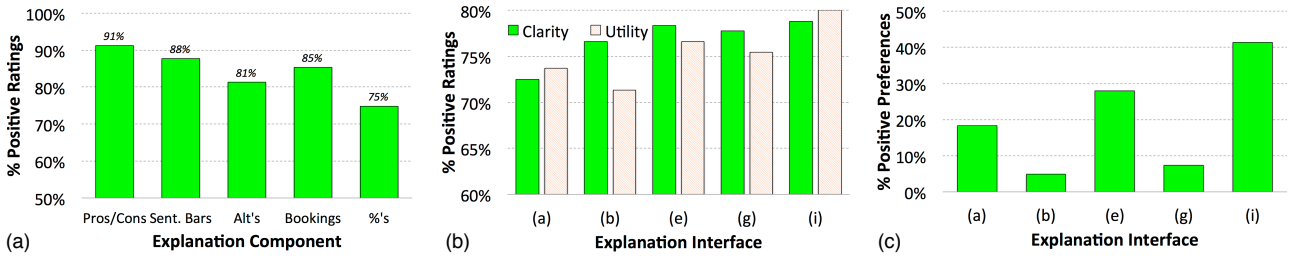

Figure 5. Results of ratings and preferences obtained from live-user study to assess the utility and clarity of different explanation interfaces

for three examples of the complete explanation interfaces. For instance, the fragment in Figure 3(e) indicates that this interface type includes sentiment bars, a reference set made up of alternative recommendations, and that percentage better/worse scores are used. An example of a complete explanation corresponding to this style is shown in Figure 4(e).

We recruited 181 participants, the majority as post-graduate researchers in a number of local universities. In each case, after some evaluation preliminaries where we explained the purpose of the study, we asked them to review the 9 different explanation interfaces and evaluate their overall clarity and utility (on a scale of 1 to 10); the presentation order was shuffled to avoid ordering effects. Based on these ratings we selected the top-5 explanation styles overall and asked the users to select one as their preferred style. After they reviewed all 9 interfaces we asked them to rate the usefulness of the different explanation components (*pros*/*cons*, sentiment bars, relationship to alternative recommendations or past bookings, and the use of percentages in better/worse scores). Next, we will describe a subset of the results obtained.

Results

To begin, Figure 5(a) shows the percentage of positive ratings (ratings of 5 or higher) received for each of the different types of explanation components. In general the ratings were very high, particularly for the use of *pros*/*cons* and sentiment bars (both with an average rating of higher that 7). Users expressed a slight preference for reference sets made up of past bookings compare to alternative recommendations, perhaps indicating that they found it easier to relate recommendations to hotels they had stayed in formerly rather than other recommendations that they had little or no knowledge about. In any event these ratings suggest that, in isolation, users found each of the various explanation components to be useful.

For reasons of space we cannot show the individual ratings scores for each of the 9 interfaces here. However, based on the ratings provided (as referred to above) we selected the top-5 most highly rated styles — variations (a), (b), (e), (g), and (i) — as shown in Figure 3. The clarity and utility scores for these top-5 versions are presented in Figure 5(b). In fact, for convenience we show the percentage of *positive ratings* (ratings of 5 or higher) in terms of clarity and utility. Note too that the explanation interface labels refer to the corresponding labels in Figure 3. We can see a preference for those styles that include richer forms of explanation data — styles (e), (g), and (i) — with clarity and utility scores in excess of 75% in each case. Explanation styles that included only the sentiment bars or the reference information – (a) and (b) — while highly rated overall did not score quite as well, especially with respect to utility. This makes sense. The participants found the various types of explanation components to be useful in their own right but even more so in combination. And in particular, the combination of sentiment bars, reference set comparisons, and percentage better/worse scores (interface (i)) achieved the highest number of positive ratings in terms of utility (80%) without any significant compromises in terms of clarity (78%).

The participants, after reviewing and rating individual interface styles, were asked to select a single favourite from the top-5 most highly rated variants. The results presented in Figure 5(c) shows the percentage of preferences that each of the top-5 most highly rated interfaces received. Again we see a strong preference for explanation style (i) which echoes its strong clarity and utility ratings above. Here, variation (i) was selected as the preferred interface by 42% of the respondents with the next best scoring variant (e) securing less than 30% of the preferences. Variant (e) was also high scoring in terms of its clarity and utility scores, but it did not rate as highly when it came to a user's single preference. Interestingly, variant (g) which scored as well as (e) and close to (i) in terms of clarity and utility ratings, did not feature prominently in the respondents final preferences; it secured less than 10% of their votes. This agrees with the preference users seem to hold for past bookings as a reference set versus alternative recommendations (see Figure 5 above); variation (g) used past bookings whilst (e) used alternative recommendations. And in the end, the combination of both reference sets in variation (i) clearly appealed to a majority of respondents.

CONCLUSIONS

We have described a novel approach to generating rich and compelling recommendation explanations mined from user-generated reviews. The results of a live-user study suggest that users found this approach to explanation to be useful, and expressed a preference for interfaces that combined a number of different explanation components. This evaluation is preliminary and limited as its focus has been solely on the explanation interface and we have not fully explored the role of these explanations in a live recommendation setting — this is a matter for future work.

ACKNOWLEDGMENTS

This work is supported by the Insight Centre for Data Analytics under grant number SFI/12/RC/2289.

REFERENCES

1. J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining Collaborative Filtering Recommendations," in *Proceedings of The 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, (Philadelphia, USA), pp. 241–250, ACM, Dec. 2000.
2. P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos, "Providing Justifications in Recommender Systems," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 38, pp. 1262–1272, Nov. 2008.
3. P. Pu and L. Chen, "Trust-Inspiring Explanation Interfaces for Recommender Systems," *Knowledge-Based Systems*, vol. 20, pp. 542–556, Aug. 2007.
4. M. Coyle and B. Smyth, "Explaining Search Results," in *Proceedings of The 19th International Joint Conference on Artificial Intelligence*, (Edinburgh, UK), pp. 1553–1555, Morgan Kaufmann Publishers Inc., July 2005.
5. G. Friedrich and M. Zanker, "A Taxonomy for Generating Explanations in Recommender Systems," *AI Magazine*, vol. 32, pp. 90–98, June 2011.
6. M. Bilgic and R. J. Mooney, "Explaining Recommendations: Satisfaction vs. Promotion," in *Proceedings of Beyond Personalization Workshop at the 2005 International Conference on Intelligent User Interfaces*, (San Diego, USA), pp. 13–18, Jan. 2005.
7. J. Vig, S. Sen, and J. Riedl, "Tagsplanations: Explaining Recommendations using Tags," in *Proceedings of The 13th International Conference on Intelligent User Interfaces*, (Florida, USA), pp. 47–56, ACM Press, Feb. 2008.
8. K. Yatani, M. Novati, A. Trusty, and K. N. Truong, "Review spotlight: a user interface for summarizing user-generated reviews using adjective-noun word pairs," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (New York, NY, USA), pp. 1541–1550, ACM, 2011.
9. J. Huang, O. Etzioni, L. Zettlemoyer, K. Clark, and C. Lee, "Revminer: An extractive interface for navigating reviews on a smartphone," in *Proceedings of the 25th annual ACM symposium on User interface software and technology*, (New York, NY, USA), pp. 3–12, ACM, 2012.
10. R. Dong, M. Schaal, M. P. O'Mahony, and B. Smyth, "Topic Extraction from Online Reviews for Classification and Recommendation," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, IJCAI '13, pp. 1310–1316, AAAI Press, 2013.
11. R. Dong, M. P. O'Mahony, and B. Smyth, "Further Experiments in Opinionated Product Recommendation," in *Proceedings of The 22nd International Conference on Case-Based Reasoning*, (Cork, Ireland), pp. 110–124, Sept. 2014.