Crowd-Based Personalized Natural Language Explanations for Recommendations

Shuo Chang F. Maxwell Harper Loren Terveen
The GroupLens Center for Social and Human-Centered Computing
Computer Science & Engineering
University of Minnesota
schang, harper, terveen@cs.umn.edu

ABSTRACT

Explanations are important for users to make decisions on whether to take recommendations. However, algorithm generated explanations can be overly simplistic and unconvincing. We believe that humans can overcome these limitations. Inspired by how people explain word-of-mouth recommendations, we designed a process, combining crowdsourcing and computation, that generates personalized natural language explanations. We modeled key topical aspects of movies, asked crowdworkers to write explanations based on quotes from online movie reviews, and personalized the explanations presented to users based on their rating history. We evaluated the explanations by surveying 220 MovieLens users, finding that compared to personalized tag-based explanations, natural language explanations: 1) contain a more appropriate amount of information, 2) earn more trust from users, and 3) make users more satisfied. This paper contributes to the research literature by describing a scalable process for generating high quality and personalized natural language explanations, improving on state-of-the-art content-based explanations, and showing the feasibility and advantages of approaches that combine human wisdom with algorithmic processes.

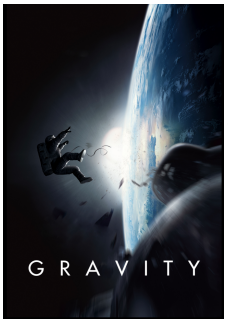
Keywords

Crowdsourcing; Recommendation Explanations; Natural Language Processing; Clustering; Word2Vec

1. INTRODUCTION

Because recommendation explanations are crucial for good user experiences with recommender systems, they have received persistent research interest [11, 20, 22]. Past research has shown several positive effects of recommendation explanations, including increasing user trust in recommender systems [18] and helping users make good decisions on recommendations [20].

Recommendation explanations on popular websites, however, are mostly generated by algorithms, hence are syntactic and formulaic. For example, "Customers who bought this item also bought ..." on Amazon and "Based on your watching history." on YouTube.



- From your MovieLens profile it seems that you prefer movies tagged as space, this movie takes you in space and it feels claustrophobic to be there. It keeps you on the edge of your seat the whole time.
- From your MovieLens profile it seems that you prefer movies tagged as visual, Gravity is unlike what we have seen on a cinema screen before and arguably it has one of the best uses of 3D in a movie.
- From your MovieLens profile it seems that you prefer movies tagged as intense, the movie a pretty intense ninety minutes, with Bullock's character constantly battling one catastrophe after another, and all of it is amazing to see.

Figure 1: Example natural language explanations for the movie "Gravity". Depending on our model of a user's interest, our system selects one of the three explanations for the user.

In comparison, people can provide effective explanations for everyday recommendations. For example, when librarians recommend books, they may summarize and highlight a book's content and help people figure out whether it is a good fit for them.

Though human effort is generally expensive, crowdsourcing enables computational processes to integrate human inputs at scale and on demand, extending the capability of current computational systems. For example, Cheng et al. [7] leverage crowd wisdom to train a machine learning system that can detect whether people in a video are lying.

Pure crowdsourcing approaches to generating natural language recommendation explanations will not succeed because most crowdworkers are not domain experts. For example, we cannot expect a crowdworker who has not seen "Gravity" to write an effective explanation for a movie fan who likes *intense* movies. Therefore, to generate natural language explanations at scale, we combine several techniques. First, we build on existing algorithmic processes that generate personalized content-based explanations [22]. Second, we combine this content model with user review text, which is popular across many recommendation systems, including IMDb and Yelp. To convert these two inputs into a set of coherent explanations that can be selected for display on a per-user basis, we rely on a combination of several crowdsourcing and algorithmic steps.

RecSys '16, September 15-19, 2016, Boston , MA, USA

DOI: http://dx.doi.org/10.1145/2959100.2959153

In this paper, we present a system that leverages this *mixed computation* – computation that combines crowdsourced data into an algorithmic processes – to generate personalized natural language recommendations at scale. See Figure 1 for three example explanations generated by our process for the movie "Gravity"; in application, we would select the one explanation for display that best matches with the current user's interests. Our system can be adapted to any recommendation application where user reviews and topic labels (e.g., categories, genres, or tags) for items are available.

We deployed the system in MovieLens and conducted a controlled experiment to evaluate the resulting explanations. Using established rubrics from past research [20], we compared the natural language explanations with tag-based explanations generated by a state-of-the-art algorithm [22].

This paper makes the following contributions:

- We created a novel system that generates personalized natural language explanations at scale with a mixed computation approach;
- Through a controlled experiment, we found that personalized natural language explanations beat a strong baseline - personalized tag explanation.

2. DESIGN SPACE AND RELATED WORK

Explanation interfaces in recommender systems communicate *why* a user might like (or dislike) a particular item. In this section we describe several key dimensions, the research and industrial work that has explored these dimensions, and where our experimental system fits.

When designing recommendation explanations, there are several "styles" [21] of explanations to choose from: case-based, collaborative-based [11], content-based [22], conversational [19], demographic-based, and knowledge-based [24]. For example, Amazon's "Customers Who Bought This Item Also Bought ..." is collaborative-based, while Pandora's "Based on what you've told us so far, we're playing this track because it features ..." is content-based.

Apart from styles of explanations, explanations can be presented in different ways, e.g., using natural language, template-based language, or a graphical representation [3, 11]. As template-based language can be generated by algorithms, it is most common on popular websites. Tintarev et al. [20] also experimented with filling a template with movie meta data. As for graphical explanations, Herlocker et al. [11], for example, experimented with histograms of ratings as explanations, finding them to be persuasive. Natural language explanations, however, can not be easily generated by an algorithm, and have received little research attention.

Recommendation explanations can be personalized or non-personalized, depending on whether or not different users see different explanations for the same item. Personalization has been shown to have positive and negative effects [20], depending on the design. Herlocker et al. [11] found that non-personalized histograms of ratings are more persuasive than personalized histograms of ratings from users' neighbors in preference space. Vig et al. [22] also observed a similar effect with tag-based explanations: personalization resulted in a better user experience in one of their designs, but had a negative effect on the other one.

Inspired by word-of-mouth recommendation explanations, we are interested in designing *personalized natural language explanations* about the content of recommended items. Our intuition is that by showing what users might like about recommended items (in a personalized fashion), we can convince them to explore or even consume unfamiliar items [12, 15]. Our goal is to help users to decide whether or not to take a recommendation. Therefore, the text should be reasonably short to avoid overwhelming users as they casually browse recommendations. We hypothesize that users have better experience with natural language, compared with template-based language.

In generating natural language explanations with human effort, we can either recruit professionals like librarians or paid crowdworkers. While professionals may generate higher quality content – though prior work has shown counter-examples [10] – crowdsourcing has the advantages of being scalable, on-demand, and relatively inexpensive.

To reduce the difficulty of writing explanations, we can show crowdworkers relevant text from user reviews, by simple text search or more advanced data mining methods [1,8, 14]. Finding crowdworkers who are familiar with items and capable of writing explanations is challenging. Showing relevant information from user reviews can address this issue: crowdworkers who have not watched a movie can still *synthesize* and *edit* already-written sentences. Current data mining methods focus on extracting *keywords and key phrases* from text. For example, Liu et al. [14] introduced an algorithm that extracts salient phrases from review data, e.g., "ice cream parlor" from Yelp's review data. These keywords offer limited help for crowdworkers. Therefore, we use simple text search to find sentences that may be relevant for explanations.

In this paper, we describe a mixed computation approach – combining crowd wisdom with algorithms. Recent research on crowdsourcing shows that such mixed computation creates exciting new applications, such as Soylent, a crowd-powered word editor [2]. More closely related, Organisciak et al. [17] proposed a crowdsourcing system to take on a collaborative filtering task - predicting user ratings. Chang et al. [6] showed that paid crowdworkers can make movie recommendations that are better perceived by users than algorithmic recommendations, and that paid crowdworkers can also produce decent quality explanations for their recommendations.

3. SYSTEM OVERVIEW

In this section, we give an overview of the system, showing how we composite three processes to generate personalized natural language explanations.

The system takes *topic labels* and *reviews* for items as input, and generates natural language explanations, which are presented to users in a personalized fashion based on their past activities, e.g., ratings, clicks, or purchases. Both topic labels and reviews are easily accessible in contemporary recommender systems. Topic labels for items can be either tags provided by users or topic entities (e.g., Google's knowledge graph entities) generated through a computational process [4, 25]. User reviews are a popular feature in e-commerce sites such as Amazon and the Apple App Store, and in specialized sites such as IMDB for movies and Yelp for restaurants.

The system consists of three processes, as shown in Figure 2 and described in more detail in Section 5:

1. *Select key topical dimensions from item topic labels.*

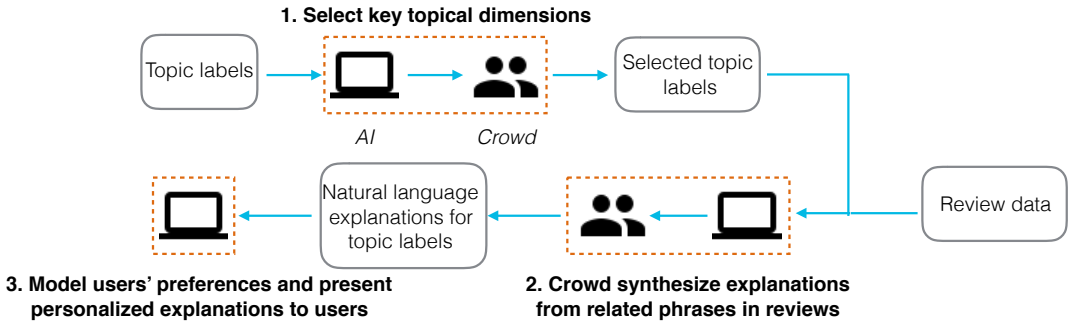


Figure 2: System overview. There are two inputs to the system - *topic labels* and *review data*. We show three procedures of the system in bold text. The first two procedures use mixed computation, with crowd and algorithm, while the last one only uses algorithm.

We first select topical dimensions with an unsupervised learning approach, and then leverage crowd wisdom to refine the output.

2. *Generate natural language explanations for the key topical dimensions.* We data mine relevant quotes for the selected dimensions, and then ask crowdworkers to synthesize the quotes into explanations.

3. *Model users' preferences and present explanations in a personalized fashion.* We compute users' interest on topical dimensions of a item based on their activities, such as rating, clicks and browsing, and then present users explanations that are most interesting to them.

4. PLATFORM

We deploy our explanation system in MovieLens, a movie recommendation website with over 3000 monthly active users. Users rate and tag movies and receive recommendations. MovieLens currently does not provide explanations for its recommendations, but this is currently the second most requested feature by its users [1].

We recruit crowdworkers from Amazon Mechanical Turk ("Mturk" for short) for the human computational tasks required in our system. To ensure quality results, we only recruited US crowdworkers who have finished more than 2000 HITs (or micro tasks) and maintain higher than 98% approval rate.

5. OVERVIEW OF SYSTEM PROCESSES

We now describe three mixed computational steps that generate explanations of the following form: "From your MovieLens profile it seems that you prefer movies tagged as *[Topical Dimension]*, *[Natural Language Explanation]*". These steps, in particular, generate two natural language components: "*[Topical Dimension]*", a personalized topic label, and "*[Natural Language Explanation]*", a crowd-synthesized explanation based on review text. For each step, we describe separately the generalizable mixed computation and the experimental implementation details.

[1]https://movielens.uservoice.com/suggestions/5585074

mafia, gangster, gangsters, mob, crime, ~~mentor~~

~~violent~~, ~~narrated~~, violence, **stylish**, ~~visceral~~, ~~stylized~~, ~~bloody~~, ~~brutality~~

masterpiece, storytelling, **drama**, dialogue

interesting, **original**

Figure 3: An example of crowd-refined tag clusters for the movie "Goodfellas". The algorithm generates four tag clusters. The crowdworkers curated the tags: a strikethrough indicates bad cluster fit, a crossout indicates inappropriateness for explanations, and red/bold indicates selection as the representative tag for the cluster.

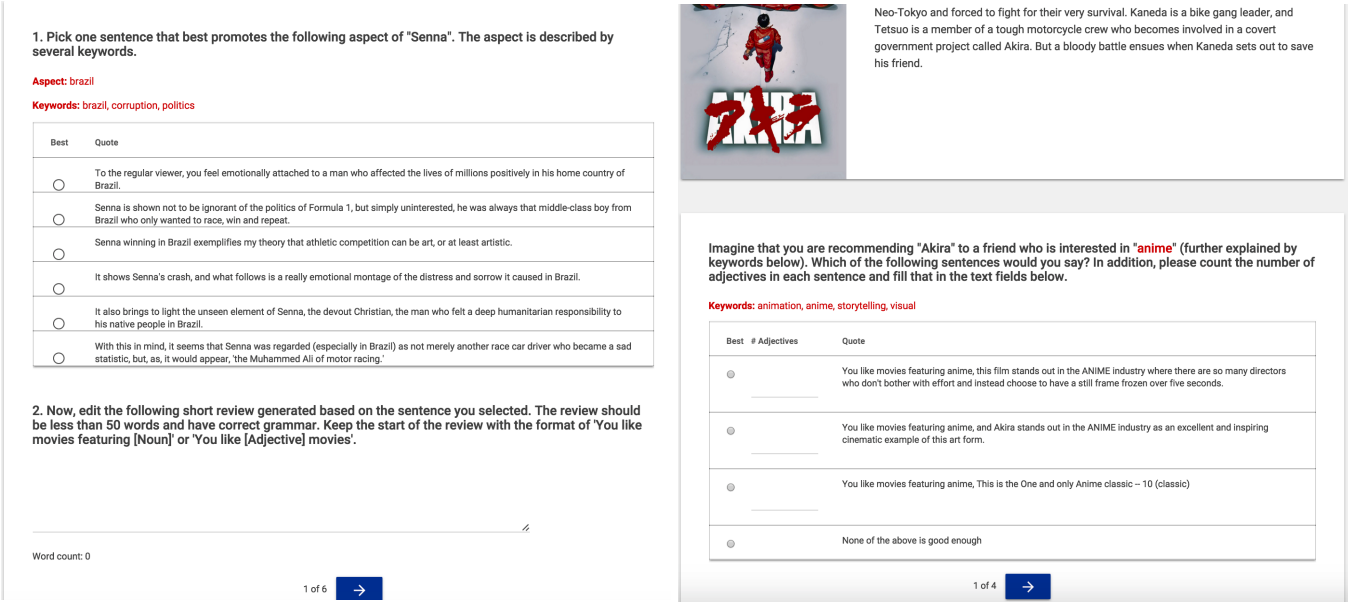
5.1 Select Key Topical Dimensions of Items

5.1.1 Description

We aim to find semantically diverse topical dimensions of an item from associated topic labels to fill the "*[Topical Dimension]*" part of the explanation. Our intuition is that to personalize an explanation for a user, we should highlight an aspect of the recommended item for which the user previously has expressed a preference. For example, some movie fans be drawn to the movie Gravity because of its space travel aspects, while others will be drawn to it for its high quality 3D effects. Therefore, we want to show different explanations (shown in Figure 1) to these two groups of users.

Using a clustering algorithm is one way to select diverse topical labels that represent a larger set of topic labels. Clustering, based on semantic similarities, groups similar topic labels together. For example, a clustering algorithm may group "3D", "visual" and "CG" together because they are semantically similar.

We leverage crowd wisdom to refine topic label clusters. Clustering, as an unsupervised learning method, is sensitive to parameter choices and notoriously difficult to evaluate. In addition, we need to find one tag to fill in "*[Topical Dimension]*", a highly subjective task. Human wisdom has been shown to be effective in improving clustering results [5]. Therefore, we decide to ask crowdworkers to refine algorithmically generated clusters and pick labels for the clusters.



(a) Write explanations

(b) Vote on best explanation

Figure 4: Interfaces used for the MapReduce work flow in the "*generate natural language explanations for the key topical dimensions*" step. The interface on the left is used in the mapper phase and interface on the right is used in the reduce phase.

5.1.2 *Implementation*

In our experimental system, we clustered most relevant tags to movies. We selected top 20 most relevant tags for each movie, using an automatic tagging system for movies - tag genome [23]. Alternatively, we could simply take top 20 most frequently applied tags for systems without tag genome. With manual inspection, we find that the 20 most relevant tags describe most key attributes of movies.

Then, we clustered, for each movie, the most relevant tags based on semantic similarities computed from movie reviews. To compute semantic similarities between tags, we first trained Word2Vec, a neural network embedding model [16], on a random sample of IMDB reviews (total size of 300MB); and then computed cosine similarities between latent vectors of tags from the embedding model. In training the word2vec model, we choose the dimension of 1000 for the embedding model, after inspecting outputs from the model with varying dimensions (ranging from 50 to 1500). With semantic similarities between tags, we constructed a similarity graph of the top 20 tags for each movie and ran an affinity propagation clustering algorithm [9] on the graph. After tuning the affinity propagation parameters, we generated between 4 to 6 tag clusters, which seemed appropriate with manual inspection, for each movie.

Lastly, we recruited crowdworkers from Mturk to refine algorithm generated tag clusters and label clusters. For each movie, we aggregated judgments, using majority voting, from three independent workers, each paid with $0.15. We instructed crowdworkers to 1) remove tags that did not belong with other tags in a cluster, 2) remove tags that were inappropriate (or offensive) to appear in the explanation template *"From your MovieLens profile it seems that you prefer movies tagged as ___"*, and 3) pick tags as labels for clusters. The example in Figure 3 shows the result of this step on the movie "GoodFellas".

5.2 Generate Natural Language Explanations for The Key Topical Dimensions

5.2.1 *Description*

Having obtained key topical dimensions for each movie in the previous process, we again use mixed computation to generate the "*[Natural Language Explanation]*" for each of the topical dimensions. As discussed in Section 2, our explanations should be short, descriptive about a certain topical dimension and well written. Given these constraints, only a human can write these explanations. However, writing explanations for a movie can be a challenging task for crowdworkers, who may have not watched the movies. Though an algorithm struggles to write explanations, it can extract relevant information about a topic dimension of a movie from user reviews, assisting crowdworkers in writing.

In addition to algorithm support, we use a two stage MapReduce [13] work flow to ensure the quality of crowd synthesized explanations. Showing algorithm-extracted text from user reviews simplifies the task of writing explanations; however, there is not a good computational method for measuring the quality of these explanations. Therefore, we had multiple independent crowdworkers synthesize explanations from selected review text in the Map phase, and then had another group of workers vote on the best explanation.

5.2.2 *Implementation*

To assist crowdworkers, we located and presented quotes describing a topical dimension of a movie. First, we found highly voted positive IMDB reviews for each movie. Then,

for each topical dimension of the movie, we searched [2] the indexed review text for sentences containing any tag in the corresponding tag clusters. Finally, we selected top 6 quotes, considering work load for crowd workers, as ranked by the number of votes and ratings on the reviews. For example, here are quotes for the aspect "**drama**, masterpiece, storytelling, dialogue" about the movie "Goodfellas":

> *As much as the true events of Henry's life have more than likely been dramatised and glamourised to a certain extent, the essence of this film IMO is that it is still a brilliantly damning portrayal of the characters and lifestyle of mobsters.*
>
> *The consistently fine acting by the large ensemble cast (both known and unknown), the cinematography, editing, dialogue, brilliant use of music, it's all breathtaking.*
>
> *The dialogue is incredible.*
>
> *Storytelling with impeccable pacing, this is what it's like when a master composer conducts his masterpiece.*
>
> *If ever the word 'masterpiece' was meant to be used, it was for this film. 'Goodfellas' is a masterpiece, pure and simple.*

In the mapper phase, we recruited three workers to synthesize explanations for each movie, paying $0.75 to each worker. As shown in the interface in Figure 4a, we had the following instructions for crowdworkers: 1) pick one quote that best describes a topical dimension (represented as tags) of a movie from 6 quotes shown; 2) rewrite the selected quote into an explanation that follows the template "From your MovieLens profile it seems that you prefer movies tagged as *[Automatically Filled in Topical Dimension]*, ___", limited to 50 words.

In the reducer phase, we recruited three workers to vote on the best explanations from the mapper phase, paying $0.40 to each worker. As shown in the interface in Figure 4b, we asked crowdworkers to vote on the best explanation for a topical dimension (represented as tags) of a given movie. For example, the resulting explanation for the aspect of "**drama**, masterpiece, storytelling, dialogue" about "Goodfellas" is:

> *From your MovieLens profile it seems that you prefer movies tagged as drama, and storytelling with impeccable pacing? Well, Goodfellas is the cinematic equivalent of a master composer conducting his masterpiece.*

5.3 Model Users' Preferences and Present Explanations in a Personalized Fashion

5.3.1 Description

Now that we have natural language explanations for multiple topical dimensions of a movie, we present users with explanations that best match with their own topic-based preferences. First, we model a user's preferences on topic labels from her activities in the recommender system, e.g., ratings and clicks. Second, we choose a natural language explanation based on the user's favorite topical dimension (represented as a topic label) for a given movie.

[2]Using Whoosh – https://pypi.python.org/pypi/Whoosh/

5.3.2 Implementation

In MovieLens, we modeled users' preferences for tags based on their movie ratings, as shown in the equation below. More formally, the tag preference of user u on tag t, denoted as $pref_{u,t}$, was computed from the set of movies M_u that user u has rated.

$$pref_{u,t} = \frac{\sum_{m \in M_u} rating_{u,m} * rel_{m,t} + K * avg_u}{\sum_{m \in M_u} rel_{m,t} + K} \quad (1)$$

where m denote a movie in M_u, $rating_{u,m}$ denotes the user's rating on movie m and $rel_{m,t}$ denotes relevance score of tag t to movie m given by algorithmic tagging system - tag genome [23]. Note that we have a user average Bayesian Prior avg_u, which is user's average preference on tags. With the choice of $K = 20$, $pref_{u,t}$ is skewed towards the prior when user has few ratings, because few ratings can not accurately represent users' preferences on tags.

When recommending a movie to a user, we picked a topical dimension, each represented with a tag, that has highest value of $pref_{u,t}$ and presented the matching natural language explanation.

6. USER EVALUATION

We evaluated our natural language explanations, compared with a baseline of personalized tag explanation, using a within-subjects user experiment in MovieLens. We pick the tag explanation designed by Vig et al. [22], because it, similar to Pandora's explanations, represents state-of-the-art content-based explanations.

We invited MovieLens users to participate in an online survey via email invitations. From January 29, 2016, we sent emails to 4000 recent active users who have more than 15 total ratings and have logged in after November 1, 2015. 220 users finished our survey (we required users to submit at least one pair of responses for inclusion) and gave 711 responses (we allowed users to continue submitting more responses after their first).

We implemented Vig et al.'s tag explanations [22] in MovieLens as follows. For a movie recommended to a user, we ranked highly relevant tags – with a relevance score higher than 4 on a 1-5 scale – based on the user's preferences, and then picked the top 5 tags in the ranking to fill in the template "*We recommend the movie because you like the following features: [tag1, ..., tag5]*" as an explanation.

6.1 Survey Design

In the survey, participants evaluated both natural language explanations and the baseline tag explanations for randomly selected movies from a set of 100 movies with high average ratings and not rated by each participant. This set of 100 movies all had an average rating of at least 3.8 on a 5 star scale; 50 were drawn from the top-500 and 50 were drawn from the 500 to 1000 most frequently rated movies. We generated natural language explanations for the 100 movies with a cost of $3.90 per movie using the previously described system. For each participant, we randomly picked an unrated movie from the set of 100 and showed it together with a baseline or experimental explanation. We asked the participant to finish at least one pair of evaluations – one baseline, one experimental – and gave them the option to evaluate more pairs.

Rubric	Statement	Stage	Metric
Efficiency	*The explanation contains right amount of information.*	2	Raw response
	I know enough about this movie to decide whether to watch it.	1, 2, 3	Change of responses before and after seeing explanation (stage 1 and 2)
	N/A	N/A	Time spent reading explanation
Effectiveness	*I am interested in watching this movie.*	1, 2, 3	Change of responses before and after watching trailer (stage 2 and 3)
Trust	*I trust the explanation.*	2	Raw response
	The explanation reflects my preferences about this movie.	2	Raw response
Satisfaction	*The explanation is easy to understand.*	2	Raw response
	The explanation is useful.	2	Raw response
	I wish MovieLens included explanations like this.	2	Raw response

Table 1: Summary of questions asked in the 3-stage user survey. We asked users to respond on 5-point Likert scale for the above statements. The "Stage" column shows the stage the statements were shown. The "Metric" column shows how we processed responses to measure the corresponding rubrics.

We evaluated explanations according to established rubrics from past research [21]. We selected the four rubrics that are most appropriate for our natural language explanations as defined below.

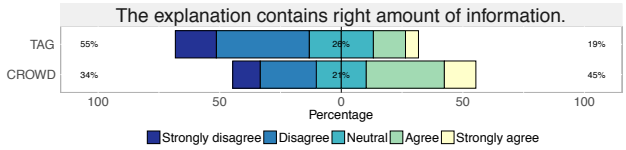
- **Efficiency:** making users acquainted with items efficiently;
- **Effectiveness:** helping users make good decisions;
- **Trust:** earning trust from users;
- **Satisfaction:** offering positive user experience and overall satisfaction.

We designed a three-stage survey to measure these four qualities of recommendation explanations, asking questions when participants 1) have only seen the title, year, genre and poster of a movie, 2) have also seen the explanation, and 3) have also watched a trailer for the movie (a preview/advertisement for the movie, typically 2-3 minutes long). In each stage, we asked participants to respond to several statements on 5-point Likert scale, from "Strongly Disagree" to "Strongly Agree". In Table 1, we summarize the questions from all three stages. Note that we asked participants to respond to two statements repeatedly in all three stages to measure the *effectiveness* and *efficiency* of explanations. More specifically, we measured *effectiveness* as the change of a participant's response to "I am interested in watching this movie" before and after watching the movie trailer. The change should be small for effective explanations, which help users to accurately gauge their interest in recommendations before consumption [21]. As asking participants to watch a full movie is unrealistic in this experimental context, we approximate this action with watching a trailer. Similarly, we measured *efficiency* with time spent reading explanations and changes in participants' responses to "I know enough about this movie to decide whether to watch it." before and after seeing explanations.

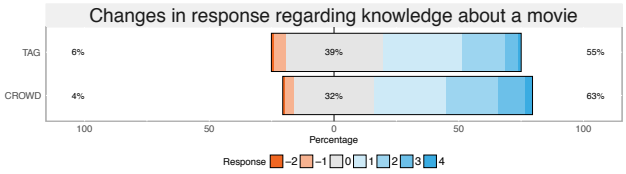
6.2 Results

In this section, we describe findings on how personalized natural language explanations compare with personalized-tag-explanation baseline in the four rubrics.

We treat raw responses to survey questions as ordinal data and use non-parametric statistical models for analysis. For *effectiveness* and *efficiency* where we measure the difference between two responses, we first map 5-point Likert responses to values 1-5, calculate the difference in values (ranging from -4 to 4) and treat the result as ordinal data. Using the Cumulative Link Mixed Models [3], we model the fixed effect of the explanation type (baseline vs. experimental) on ordinal responses, with user and movie as random effects. We include the two random effects to capture the variances of responses across users and variances caused by different movies shown in the survey.



(a)



(b)

Figure 5: Survey responses to questions regarding *efficiency*. Natural language explanations - labeled with "crowd" - contained a more appropriate amount of information (a) and helped subjects more with decision-making (b).

6.2.1 *Efficiency*

As compared with our baseline explanations, participants perceive the natural language explanations to have more information and a more appropriate amount of information; on the other hand, they take longer to read. Participants in 45% of cases agreed with the statement "The explanation contains right amount of information." for natural language explanation, compared to only 19% for tag explanations (CLMM, $N = 711$, $p \sim 0$), as shown in Figure 5a. More specifically, natural language explanations contain more information than tag explanations, as evident by changes of responses on "I know enough about this movie to decide whether to watch it.". We observe greater changes towards positive direction when showing natural language explana-

[3]Included in 'ordinal' package of R

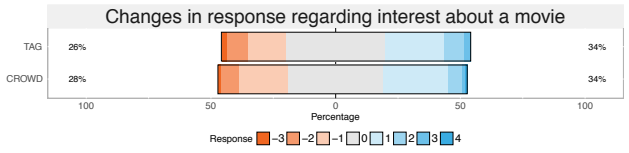


Figure 6: Difference in interest before and after watching the trailer to measure *effectiveness*;no statistically significant difference.

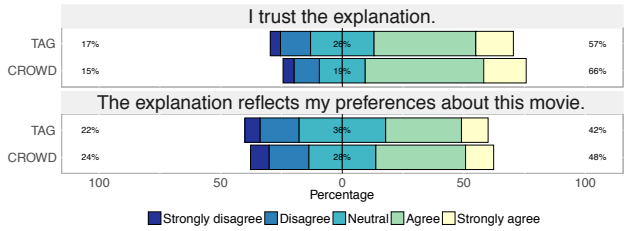


Figure 7: Survey responses to questions regarding *trust*. Participants trusted natural language explanations more than tag explanations.

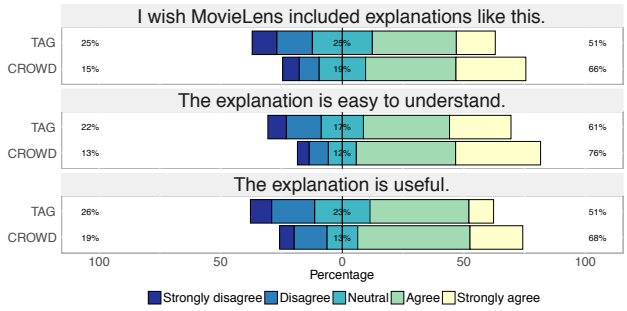


Figure 8: Survey responses to questions regarding *satisfaction*. Across the three questions, participants gave more positive responses for natural language explanations as compared with tag explanations.

tions than tag explanations (CLMM, $N = 711$, $p < 0.01$) as shown in Figure 5b. This is as expected, because natural language explanations are longer and more descriptive than tag explanations. As a result, we observe the increased time participants spent reading natural language explanations as compared with tag explanations: there is a 0.38 difference of log transformed seconds spent on reading.

6.2.2 *Effectiveness*

We observe that natural language explanations are no different than tag explanations in terms of effectiveness, which is measured as the changes of responses on "I am interested in watching this movie." when seeing explanations and having watched trailers. As shown in Figure 6, a participant's interest level, in fact, is slightly more likely to decrease for natural language explanations than for tag explanations (28% vs. 26%). This can be attributed to the fact that natural language explanations are more persuasive, hence participants may became overly excited after reading the explanations. Note that in both conditions, 1/3 of the users reported the same interest level after watching trailers; given that trailers are usually rich with information that assists in decision-making, this result is encouraging regarding the overall effectiveness of both types of explanations.

6.2.3 *Trust*

We find that participants trusted personalized natural language explanations more than the baseline tag explanations as shown in Figure 7. For the statement "I trust the explanation.", participants were significantly (CLMM, $N = 711$ [4], $p < 0.05$) more likely to agree when seeing natural language explanations than seeing tag explanations (66% vs. 57% agree). We find more agreement on the statement "The explanation reflects my preferences about this movie." for natural language explanations, but this effect is small and not statistically significant.

6.2.4 *Satisfaction*

Participants rated natural language explanations more highly on all three satisfaction questions, shown in Figure 8 (CLMM, $N = 711$, $p < 0.001$ for all three): 68% of responses agreed that natural language explanations are "useful" compared to 51% for tag explanations. Though natural language explanations are longer and contain more information, 76% of the responses agreed that they are "easy to understand" compared to 61% of the responses for tag explanations. Overall, 68% of responses wished to include natural language explanations in MovieLens compared to 51% for tag explanations.

[4]220 participants have options to evaluate more than 1 pair

7. DISCUSSION

The user experiment shows that natural language explanations generated from our process are better received by users compared to the personalized-tag-explanation baseline, which represents state-of-the-art content-based explanation. Compared with the baseline, users perceived that natural language explanations are *more trustworthy*, *contain a more appropriate amount of information* and *offer a better user experience*. To our surprise, though, natural language explanations are comparable in effectiveness for helping users make decisions on whether to take recommendations. Considering that users only spend a small amount of effort reading explanations, both content-based explanations are fairly effective in helping users make good decisions. However, apart from the content of items, many other factors may affect users' decision on whether to take recommendations, such as their trust in the system, social influence and past experience [12]. This is perhaps the reason why users perceive that natural language explanations contain richer information but are not significantly more effective in decision-making support.

Though natural language explanations written by humans are superior to automatically generated explanations, human labor is expensive and difficult to scale up. Our system addresses this challenge through a mixed computation model that combines intelligent algorithms with crowdsourced inputs. Crowdsourcing allows our system to quickly recruit large number of workers as needed. Our algorithmic processes model the content of items and extract quotes from existing user reviews, two steps that effectively reduce the

human effort required per movie. In our experiment we generated natural language explanation for 100 movies at the cost of $3.90 per movie in a short amount of time. For organizations that can afford this per-item cost, we believe our approach could be scaled to a much larger item space.

Our explanation process underscores the potential of mixed computation approaches that combine algorithms with crowd wisdom. Recent research in machine learning, especially deep neural networks, has advanced the limit of computers in image/voice recognition tasks and natural language conversations – advances that are made possible with human-generated training data. For tasks that remain too challenging for a purely algorithm solution (e.g., writing recommendation explanations) crowdsourcing may be a scalable solution. In the long run, crowd-generated data can be fed into machine learning processes to enable the automation of human work, pushing the limit of algorithms.

8. CONCLUSIONS

This paper makes the following contributions:

- We designed and implemented a novel scalable system that generates personalized natural language explanations with a mixed computation approach, which combines crowdsourcing and algorithms.
- Through a controlled experiment with 220 MovieLens users, we find that, compared with personalized tag-based explanations, personalized natural language explanations 1) contain a more appropriate amount of information, 2) earn more trust from users, and 3) make users more satisfied.

9. ACKNOWLEDGMENT

We thank volunteers from MovieLens community and anonymous workers on Mturk. We also thank NSF for funding this research with grant 1017697, 964695 and 1111201.

10. REFERENCES

[1] S. Bedathur, K. Berberich, J. Dittrich, N. Mamoulis, and G. Weikum. Interesting-phrase mining for ad-hoc text analytics. *Proceedings of the VLDB Endowment*, 3(1-2):1348–1357, 2010.

[2] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich. Soylent: a word processor with a crowd inside. In *UIST*, 2010.

[3] S. Bostandjiev, J. O'Donovan, and T. Höllerer. TasteWeights: a visual interactive hybrid recommender system. In *RecSys*, 2012.

[4] S. Chang, P. Dai, J. Chen, and E. H. Chi. Got Many Labels?: Deriving Topic Labels from Multiple Sources for Social Media Posts using Crowdsourcing and Ensemble Learning. In *WWW*, 2015.

[5] S. Chang, P. Dai, L. Hong, S. Cheng, Z. Tianjiao, and C. Ed. AppGrouper: Knowledge-based Interactive Clustering Tool for App Search Results. In *IUI*, 2016.

[6] S. Chang, M. F. Harper, L. He, and L. G. Terveen. CrowdLens: Experimenting with Crowd-Powered Recommendation and Explanation . In *ICWSM*, 2016.

[7] J. Cheng and M. S. Bernstein. Flock: Hybrid crowd-machine learning classifiers. In *CSCW*, 2015.

[8] A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han. Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment*, 8(3):305–316, 2014.

[9] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[10] F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan. Predictors of answer quality in online Q&A sites. In *CHI*, 2008.

[11] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *CSCW*, 2000.

[12] A. Jameson, M. C. Willemsen, A. Felfernig, M. Gemmis, P. Lops, G. Semeraro, and L. Chen. *Recommender Systems Handbook*, chapter Human Decision Making and Recommender Systems, pages 611–648. Springer US, Boston, MA, 2015.

[13] A. Kittur, B. Smus, S. Khamkar, and R. Kraut. Crowdforge: Crowdsourcing complex work. *UIST*, 2011.

[14] J. Liu, J. Shang, C. Wang, X. Ren, and J. Han. Mining quality phrases from massive text corpora. 2015.

[15] S. M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough. In *CHI EA*, 2006.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[17] P. Organisciak, J. Teevan, S. Dumais, R. C. Miller, and A. T. Kalai. A Crowd of Your Own: Crowdsourcing for On-Demand Personalization. In *HCOMP*, 2014.

[18] R. Sinha and K. Swearingen. The role of transparency in recommender systems. In *CHI EA*, 2002.

[19] C. A. Thompson, M. H. Göker, and P. Langley. A personalized system for conversational recommendations. *Journal of Artificial Intelligence Research*, 21(1):393–428, Mar. 2004.

[20] N. Tintarev and J. Masthoff. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, 22:399–439, 2012.

[21] N. Tintarev and J. Masthoff. *Recommender Systems Handbook*, chapter Explaining Recommendations: Design and Evaluation, pages 353–382. Springer US, Boston, MA, 2015.

[22] J. Vig, S. Sen, and J. Riedl. Tagsplanations. In *IUI*, 2008.

[23] J. Vig, S. Sen, and J. Riedl. The Tag Genome. *ACM Trans. on Interactive Intelligent Systems*, 2(3):1–44, sep 2012.

[24] W. Wang and I. Benbasat. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *J. Manage. Inf. Syst.*, 23(4):217–246, May 2007.

[25] S.-H. Yang, A. Kolcz, A. Schlaikjer, and P. Gupta. Large-scale high-precision topic modeling on twitter. In *KDD*, Aug 2014.