World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:18, No:1, 2024

Impact of Similarity Ratings on Human Judgement

Ian A. McCulloh, Madelaine Zinser, Jesse Patsolic, Michael Ramos

***Abstract*—**Recommender systems are a common artificial intelligence (AI) application. For any given input, a search system will return a rank-ordered list of similar items. As users review returned items, they must decide when to halt the search and either revise search terms or conclude their requirement is novel with no similar items in the database. We present a statistically designed experiment that investigates the impact of similarity ratings on human judgement to conclude a search item is novel and halt the search. In the study, 450 participants were recruited from Amazon Mechanical Turk to render judgement across 12 decision tasks. We find the inclusion of ratings increases the human perception that items are novel. Percent similarity increases novelty discernment when compared with star-rated similarity or the absence of a rating. Ratings reduce the time to decide and improve decision confidence. This suggests that the inclusion of similarity ratings can aid human decision-makers in knowledge search tasks.

***Keywords*—**Ratings, rankings, crowdsourcing, empirical studies, user studies, similarity measures, human-centered computing, novelty in information retrieval.

I. Introduction

RECOMMENDER systems may involve entering some input, such as a search term, document, image, or other content and the AI system will return a list of similar items [1]. These systems are particularly useful when there exists a large corpus of potentially similar items. Items are typically presented in order of their relevance to the search term or their similarity, where most similar items are presented first. One challenge for users working with a recommender system is to determine when a particular search has been exhausted and should be abandoned. Some systems attempt to aid the user by including a similarity rating such as a rating scale of one to five stars or a percentage similarity. We look at how these differences impact human assessment or decision making.

We present a statistically designed experiment to address the following five research questions:

RQ1. Does the inclusion of similarity ratings impact human judgement in knowledge search tasks?

RQ2. Does the inclusion of similarity ratings increase the likelihood to conclude an item is novel?

RQ3. Does the clustering of item similarity moderate judgement differences between ratings and rankings?

RQ4. Does the inclusion of similarity ratings decrease the time to decide an item is similar/novel?

RQ5. Does the inclusion of similarity ratings impact a user's decision confidence when they assess similarity/ novelty?

II. Background

There exists a great deal of literature on recommender systems and the impact ratings have on the performance of those systems, albeit with mixed results. A common thread throughout the literature, however, is that the choice of the particular rating scale used has an impact on human judgement [2]-[6]. Cummins and Gullone [6] argue that human judgement in response to a rating scale is more a matter of psychology than mathematics or statistics. Given the existence of human subjectivity in the perception of ratings, the user/system design will likely affect bias, consistency, and confidence in judgement tasks.

Cosley et al. [7] investigated recommender systems that involved opinion data and concluded that different rating scales were well correlated and did not affect performance. Vaz et al. [8] investigated recommender systems for books to improve library useability. They found that rating scales with smaller granularity, such as 0-5 stars as opposed to 0-100 percentage, will achieve a lower mean absolute error in a rating prediction task and can thus be considered a better result. In their application, they also found that users would manually rate more content with smaller granular scales and thus increase data volume to improve search performance. From these experiments and others, it is unclear whether differences in rating scale correlations are affected by the knowledge domain, the user-rating engagement, or the rating scale. In other studies, when users rate the same item on different scales, as much as 40% of ratings are considerably different indicating that human judgment may be impacted by the choice of scale [9], [10].

Much of the differences across types of rating scales impact different aspects of what might be considered a good quality in a scale, such as reliability, validity, time-to-decide, discernment, and accuracy. An examination of Likert-scales with 2-11 points and a 101-point scale, revealed that scales with less than 5-points were the least reliable, least valid and least discerning, but demonstrated a shorter time-to-decide [11], [12]. Participants preferred more granular scales. Another study involved participants rating sets of movies that had already been rated on a 5-point scale with either a binary thumbs up/down, a 6-point no-zero scale ranging from -3 to +3, and a half-star scale ranging from 0.5 to 5. The authors found that all scales were well correlated, but that users tended toward higher mean ratings on the binary scale than the original five-point scale [7]. These findings are consistent with Preston and Colman [11] who concluded that scales with high granularity require more thought, while low granularity scales make extreme ratings more likely. Tradeoff between high and low granularity scales

Ian A. McCulloh and Madelaine Zinser are with Johns Hopkins University, 11100 Johns Hopkins Rd, Laurel, Md 20753 (e-mail: Imccull4@Jhu.Edu, Mzinser1@Jhu.Edu).

Jesse Patsolic and Michael Ramos are with Accenture, 1201 NW New York Ave, Washington Dc 20005 (e-mail: Jesse.L.Patsolic@Accenture.Com, Michael.Ramos@Accenture.Com).

may therefore depend on the desired quality objectives of reliability, validity, and discernment on one hand versus a shorter time-to-decide, and volume of ratings on the other hand.

All of the literature presented investigates the judgement of a human rating items in a recommender system. It does not directly assess human judgement of results returned by a recommender system such as the point at which a user concludes a search is complete; or the conclusion whether a search item is returning many similar returns or that the item is novel. This paper will extend our understanding of human judgement and recommender systems to human appraisal of returned results. The specific use-case is searching to see if a given item, the search item, is similar to others in a database or whether it is novel.

III. Method

We conduct a three factor, factorially designed experiment to assess human judgements of similarity. For this, 450 participants were recruited via Amazon Mechanical Turk and paid $0.10 to review the results of four knowledge search tasks and answer seven questions. Responses were captured using Qualtrics in an integrated user interface designed for this experiment.

The stimulus material consisted of displaying a recent movie title (the search item) and the 10 most similar movie titles as results from a search engine. The movies were selected from a Google search of "most popular movies in 2022" accessed on 3 October 2022. The movie titles were passed to a similarity search engine, Best Similar [13], to obtain similar movie titles, plot synopsis, cover art, and similarity score. The final movies used in the experiment were selected such that they met the criteria laid out in the statistical experimental design described below. Participants were able to review the title, year of release, cover art, and plot synopsis for the search item movie and the top 10 most similar movies returned by the search engine. Instructions to the participant included an example task to ensure understanding and were stated as follows:

> This Amazon Mechanical Turk task will require you to look at plot synopses of four movies and the synopses of the 10 most similar movies for each, as assessed by an artificial intelligence search engine. You will then answer seven questions on a scale of 1 to 5 related to the likely success of the movie. The 10 similar movies will be ranked from one to 10 in order of similarity. For some of the movies a percentage similarity score will be included; for some a five-star similarity rating will be included; and for some, no rating is included.

For example, Thor: Love and Thunder (2022) is the movie. The 10 most similar movies include: 1) Doctor Strange in the Multiverse of Madness (2022) with 100% similarity, 2) Thor (2011) with a 100% similarity, … 10) Shang-Chi and the Legend of the Ten Rings (2021) with a 24% similarity.

For each task, the participant was asked seven questions, including one yes/no and six 5-point Likert-scale responses:

1. How unique is this movie? [5 point scale: very unique, somewhat unique, neutral, somewhat common, very common]
2. How similar is the plot to those of the listed movies? [5 point scale: very similar, somewhat similar, neutral, somewhat dissimilar, very dissimilar]
3. How likely is this movie to be grouped on Netflix with those listed? [5 point scale: very likely, somewhat likely, neutral, somewhat unlikely, very unlikely]
4. Will this movie attract the same audience as those listed? [5 point scale: very likely, somewhat likely, neutral, somewhat unlikely, very unlikely]
5. How likely would you expect box office sales for this movie to be equivalent to those of the movies listed? [5 point scale: very likely, somewhat likely, neutral, somewhat unlikely, very unlikely]
6. Have you seen or do you plan to see this movie? [yes/no]
7. How confident are you in your answers? [5 point scale: very confident, somewhat confident, neutral, unsure, very unsure]

The first five questions are items for a composite scale and meant to assess human judgement of novelty. The composite scale is assessed for internal consistency using Cronbach's alpha. Item one is reverse coded. The composite novelty scale is an average of the items, such that a score of 1 is an assessment of similarity and a score of 5 is an assessment of novelty. Question six is intended to assess whether participant familiarity with the movie genre may moderate their judgement. Question seven will assess decision confidence. Time-to-decide is measured on the back end through the Qualtrics system.

Three factors were systematically varied in a statistically designed experiment to assess the five research questions and displayed in Table I. The first factor is Rating which consists of three categorical levels, Percent, Star, Rank. Under the percent condition, participants are shown the percentage similarity score returned by the BestSimilar search engine. Under the star condition, the percentage similarity is converted into a five-star similarity rating such that 5 stars = 90-100%; 4 stars = 70-89%; 3 stars = 50-69%; 2 stars = 30-49%; and 1 star = 0-29%. Under the rank condition, similarity ratings are omitted, and participants only see the returned items in order of similarity, but no objective rating information.

The second factor is clustering and has two levels. It is possible that a search returns a set of responses such that one subset is very similar, and another subset is very dissimilar, or the scores may more gradually vary from similar to dissimilar. This might moderate judgement on similarity. The mean squared error (MSE) is calculated for all similarity scores of the movies returned by the search engine. The scores are then clustered into two groups, high and low similarity. A clustered (+) condition occurs when the MSE of the percent similarity scores are reduced by more than 70% in a clustered condition. A non-clustered (-) condition occurs when the MSE is reduced by less than 20%. No movies were selected that would have a moderate MSE reduction between 20-70%.

The third factor is similarity. This factor measures how similar the ratings are to each other. To vary between two levels in both the clustered and non-clustered condition, we measure how similar the ratings are to each other within their clusters. In the case of the non-clustered condition, this is simply a single cluster. A similar (+) condition occurs when the within cluster sums of MSE of the percent similarity scores are less than 120. A dissimilar (-) condition occurs when within cluster sums of MSE of the percent similarity scores are greater than 225. The factor score ranges from 0.0 to 540.2.

TABLE I
THREE-FACTOR, FACTORIAL DESIGNED EXPERIMENT

Exp ID	Clustering	Similarity	Rating
1	(-)	(-) 314.4	Star
2	(+) 75.8%	(-) 464.4	Star
3	(-)	(+) 115.6	Star
4	(+) 99.7%	(+) 19.9	Star
5	(-)	(-) 328.4	%
6	(+) 74.9%	(-) 461.9	%
7	(-)	(+) 0.0	%
8	(+) 86.6%	(+) 84.5	%
9	(-)	(-) 540.4	Rank
10	(+) 73.0%	(-) 227.4	Rank
11	(-)	(+) 28.9	Rank
12	(+) 88.0%	(+) 108.3	Rank

The 12 movies selected to meet the factorial designed experiment requirements are displayed in Table II.

TABLE II
PARTICIPANT STIMULUS MOVIES

Exp ID	Movie
1	Good Luck to You
2	The Happening
3	Hustle
4	Top Gun: Maverick
5	Elvis
6	After Yang
7	Scream
8	Turning Red
9	Kimi
10	Cyrano
11	Downton Abby: A New Era
12	The Outfit

It should be noted that in analysis, the two factors clustered, and similarity can and will be treated as either numeric variables or as categorical indicator variables to assess whether there is any impact on response variables.

IV. RESULTS

Data were collected from 450 participants with response duration ranging from 2-11 minutes per assessment and a mean decision time of 5.2 minutes and standard deviation of 1.86 minutes.

The five Likert-scale responses for assessing novelty/similarity were not found to be internally consistent, with a Cronbach's alpha of 0.49. This was due to the reverse-coded question "how unique…" when the other responses were "how similar…". Dropping the first response and using the other four items results in an internally consistent scale with Cronbach's alpha of 0.79. The revised composite scale is used for the remainder of the analysis.

RQ1 was supported. There is a difference between ratings and rankings on human decision making. There is an impact on judgement (e.g., assessing whether an item is novel) (F = 30.91, p < 0.0001). There is an impact on decision confidence (F = 22.07, p < 0.0001). There is a reduction in the time required to decide.

RQ2 was supported as illustrated in Fig. 1 and further supported in Table III. A participant was more likely to assess that an item was novel when either percentage or star ratings were provided. Fig. 1 shows the participants' assessment of novelty along the y-axis, such that the higher the score, the more novel the item is assessed. The x-axis shows the ground-truth similarity as measured by the MSE of percentage rating from the source data as described above. The negatively sloped lines show that items with more similar movie returns are assessed lower on the novelty scale. While slopes for rank-only and star rating are similar, increased slope is observed for percentage ratings which may indicate that percentage ratings offer greater discernment over judgements of novelty than star ratings.

RQ3 was not supported. The regression results suggest that clustering of similarity scores does not impact judgement (T = -1.15, p = 0.2519). This finding is robust under different model term selections and for treating the clustering factor as a dichotomous clustered versus unclustered condition as well as the continuous numeric MSE reduction condition.

RQ4 was supported. Ratings reduces time to decide. The percent rating reduces decision time by 44 seconds on average, which is a 14% reduction (T = 2.327, p = 0.0201). The star rating achieves a slightly greater reduction in decision time of 52 seconds, which is a 17% reduction (T = 2.763, p = 0.0058) and consistent with prior findings [11], [12].

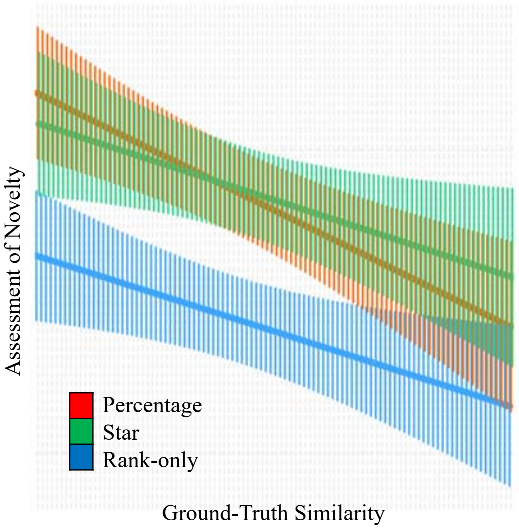

Fig. 1 Ground-truth similarity vs. participant assessments of novelty

RQ5 was supported. Ratings improve decision confidence. Recall that the decision confidence was measured on a five-point scale where a score of one was very confident and a score of five was very unsure. The percent rating improves confidence by 0.11 points, which is an 8% improvement in average confidence score (T = 2.485, p = 0.0131). The star rating improves confidence by 0.13 points, which is a 10% improvement in average confidence score (T = 3.148, p = 0.0017).

TABLE III
REGRESSION RESULTS FOR ASSESSMENT OF NOVELTY AS A FUNCTION OF RATINGS, CLUSTERING, SIMILARITY, AND FAMILIARITY

Coefficients	Estimate	Std. Err	T-val	p-val
Intercept	4.6280	0.0649	71.37	<0.001
Percentage	0.1340	0.0393	3.14	<0.001
Star	0.1359	0.0396	3.43	<0.001
Clustering	-0.0004	0.0004	-1.15	0.2519
Similarity	-0.0004	0.0001	-4.96	<0.001
Familiarity	-0.4375	0.0434	-10.12	<0.001

F-statistic = 30.91 on 5 and 1842 df, p-val < 0.0001

V. RESULTS

Human judgements of novelty are found to be affected by the presence of similarity ratings when compared to rank-ordered lists without ratings. Determinations of similarity/novelty are subjective, however. We find that when human judges are presented with ratings, either percentage similarity or a star-rating system, they are more likely to conclude an item is novel. Whether this is a desirable outcome is unclear. For example, if this study generalized to patent examiners reviewing patent applications against similarly filed patents, the use of ratings would likely increase the number of approved patent applications when compared to using rank-ordered content alone. Whether this is a good outcome is dependent upon the threshold and sensitivity for the definition of what is novel. Certainly, from the perspective of the patent applicant, this is a positive outcome. The patent office, however, may or may not wish to approve more patents. Other outcomes from the use of ratings, however, are uniformly positive.

Findings suggest that the use of percentage ratings may improve decision discernment. When an item is more novel, the reviewer is more likely to offer a judgement of novelty and when an item is more similar or common, the reviewer is less likely to offer a judgement of novelty. Discernment is a positive outcome that favors the use of percentage ratings.

Ratings also decrease the time to decide, allowing people to render their judgement faster. While several seconds saved in rendering a single decision may be trivial, if applied to an organization with several thousand people employed in knowledge search tasks, making many queries each day, a 14-17% reduction in labor costs is significant. This time saving can either be reinvested in backlog reduction or the same work can be completed with less labor, achieving fiscal impact.

Ratings also improved decision confidence, making people feel more comfortable with their decisions. While this may complicate the issue of determining whether a judgement of novelty is correct, as in the difference between ratings and rank-only, we observe that the judge is more confident of their decision when presented with similarity ratings. As in the case of time-to-decide, star ratings improve decision confidence over percentage ratings.

These findings offer several considerations for the design of any recommender system. Rank-only presentation of results should be used if the goal is to bias users towards decisions of similarity (non-novelty), but comes at the cost of decision time and confidence. Percentage ratings offer the best judgement discernment and improved decision time and confidence. Star ratings lack the benefit of improved discernment that percentage ratings demonstrate, however, star ratings offer a 20-25% further improvement for decision time and confidence. Given the subjective nature of a novelty decision in the first place, we feel that the use of similarity ratings is an improvement over rank-only presentation of results in almost all applications.

There are several limitations with this experiment. Prior research has suggested that human judgement of rating scales is domain dependent. While these findings are consistent with Cosley et al. [7], also in the movie domain, they may not generalize to other contexts, such as library book search, or patent application review. The research methodology proposed in this paper, however, may inform a more efficient and parsimonious experiment to validate conclusions for other domains. This paper presents one of the few studies that directly investigate human judgement for results returned by an objective and well-validated AI/ML recommender system, while others investigate the consistency of human-rated content. It provides important insight into human judgement under these conditions, but should be further supported by future studies before any strong conclusions are drawn. Participants were drawn from an online crowdsourcing pool and their commitment to the research may be less than those in other studies. This risk was mitigated by the use of an assessment of response time, and internal consistency validation of participant responses.

This paper presents empirical insight into tradeoffs in the use of different rating scales for recommender systems. It also presents a repeatable experimental design and framework to evaluate the use of different rating scales across varying knowledge domains. These findings can improve the user experience of recommender systems designed for knowledge search. It is the opinion of the authors that a similarity rating is generally preferred over rank-only and that a rating scale should consist of a minimum of five points or higher. When greater discernment is the primary objective, a percent similarity rating is preferred. When shorter time-to-decide and improved decision confidence is the primary objective a star similarity rating is preferred.

REFERENCES

[1] Melville, P., Sindhwani, V., Sammut, C. and Webb, G.I., 2010. Recommender Systems. In Encyclopedia of machine learning.

[2] Cena, F., Gena, C., Grillo, P., Kuflik, T., Vernero, F. and Wecker, A.J., 2017. How scales influence user rating behaviour in recommender systems. *Behaviour & Information Technology*, *36*(10), pp.985-1004.

[3] Garland, R. 1991. “The Mid-Point on a Rating Scale: Is it Desirable.” Marketing Bulletin 2: 66–70.

[4] Friedman, H. H., and T. Amoo. 1999. “Rating the Rating Scales.” Journal of Marketing Management 9 (3): 114–123.

[5] Amoo, T., and H. H. Friedman. 2001. “Do Numeric Values Influence Subjects Responses to Rating Scales?” Journal of International Marketing and Marketing Research 26: 41–46.

[6] Cummins, R., and E. Gullone. 2000. “Why We Should Not Use 5-Point Likert Scales: The Case for Subjective Quality of Life Measurement.” In

Proceedings of the Second International Conference on Quality of Life in Cities, 74–93. Singapore: The School.

[7] Cosley, D., S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl. 2003. "Is Seeing Believing?: How Recommender System Interfaces Affect Users" Opinions." In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '03, 585–592. New York: ACM.

[8] Vaz, P. C., R. Ribeiro, and D. M. de Matos. 2013. "Understanding the Temporal Dynamics of Recommendations Across Different Rating Scales." UMAP Workshops 2013, CEUR-WS.org.

[9] Cena, F., F. Vernero, and C. Gena. 2010. "Towards a Customization of Rating Scales in Adaptive Systems." In User Modeling, Adaptation, and Personalization – 18th International Conference, UMAP 2010, edited by P. De Bra, A. Kobsa, and D. N. Chin, Big Island, HI, USA, June 20–24. Proceedings, Volume 6075 of Lecture Notes in Computer Science, 369–374.

[10] Gena, C., R. Brogi, F. Cena, and F. Vernero. 2011. "The Impact of Rating Scales on User's Rating Behavior." In User Modeling, Adaption and Personalization – 19th International Conference, UMAP 2011, Girona, Spain, July 11–15. Proceedings, Lecture Notes in Computer Science 6787, 123–134.

[11] Preston, C., and A. Colman. 2000. "Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences." Acta Psychologica 104 (1): 1–15.

[12] Weng, L. J. 2004. "Impact of the Number of Response Categories and Anchor Labels on Coefficient Alpha and Test-Retest Reliability." Educational and Psychological Measurement 64 (6): 956–972

[13] Best Movies, https://bestsimilar.com/movies, retrieved on October 31, 2022.