Reproduction Study - How do Gain and Discount FunctionsAffect the Correlation between DCG and User Satisfaction?

Alexander Telenkov
e1467735@student.tuwien.ac.at
Technische Universität Wien
Vienna, Austria

Matteo Rossi Reich
e12006891@student.tuwien.ac.at
Technische Universität Wien
Vienna, Austria

Maximilian Michel
e1552754@student.tuwien.ac.at
Technische Universität Wien
Vienna, Austria

ABSTRACT

The main focus of this analysis is to reproduce the experimental setup, experiments, and results as explained in a study by Urbano Marrero [1] from ECIR 2015 and to thereby prove the numbers and findings reported in the paper. In the process the method is checked for irregularities or contradictions which could consequently call into question conclusions that have been made. The selected experiment reflects an empirical analysis of the effect that the gain and discount functions have on the correlation between Discounted Cumulative Gain (DCG) and user satisfaction. The authors come to the conclusion, that a combination of linear gain and constant discount shows the best correlation with user satisfaction. The original paper provides a link to a Github repository[1] of functions used in the calculations of this result. This serves as a baseline for this project in assessing the reproducibility and the possibility to expand on the findings of the initial work . We find that the results are reproducible and that there are no irregularities that might undermine the original conclusion but fall short of rerunning the experiment with new data due to insufficient funds limiting the access to the crowd-sourcing method used in the original research.

KEYWORDS

reproducibility, DCG, User Satisfaction, Gain, Discount, Data Sets, User Preference

1 INTRODUCTION

The paper How do Gain and Discount Functions Affect the Correlation between DCG and User Satisfaction? written by Julian Urbano and Monica Marrero [1] reflects experimental results that show the connection between user satisfaction and DCG score. To be more precise, results are related "to the probabilities that users find a system satisfactory given a DCG score, and that they agree with a difference in DCG as to which of two systems is more satisfactory"[1]. The relationship of 36 different combinations of gain and discount has been analyzed. It has been found that a linear gain and a constant discount have the best correlation with user satisfaction. Furthermore, the researchers discovered an approach that assists in studying the connection between effectiveness measures and user satisfaction. Hence, this method has been applied for a music recommendation task with DCG and a range of gain and discount functions. Moreover, the study presents facts to confirm that the usual choice of exponential gain can underestimate user satisfaction. Additionally, the research indicates that all types of discount tend to underestimate user satisfaction as well showing that the ranking does not impact user's choice and decision.

Our ultimate goal is either to prove the numbers, findings and results as exact as possible, or alternatively, to uncover irregularities as well as contradictions and consequently call into question conclusions that have been made by Urbano Marrero[1].

As a first step of this paper we have analyzed the initial study in order to understand the main idea of the research and to make clear what the authors wanted to demonstrate to their readers. Next, we made ourselves familiar with the code available in the Github repository and finally we focused on the reproduction of the results from the paper.

2 EVALUATION METRICS

In the original paper [1] the authors showed a new way to explore the relationships between effectiveness measures and user satisfaction. They conducted a user study where different recommendation systems were analyzed for their influence on user satisfaction. Each subject was presented with two lists of results resulting from a query and had to assess which of two systems they preferred and individually if they are good or bad . The resulting values were used to compute a P (Sat|φ) and a P (Pref |$\Delta\varphi$) score for each system. These were then compared to scores for all 36 corresponding DCG formulations. Based on the analysis of mapping for 6 gain and 6 discount functions we can see that the usual exponential gain underestimates satisfaction. Moreover, all other forms of discount function by the same principle.

3 EXPERIMENTAL SETUP

Urbano Marrero have provided the link to a public repository on GitHub ensuring the availability of codes and datasets that have been processed. In order to make the implementation process easier instructions how to run the code with the specific data set have been given as well. The software requirements of the project are limited to the R Base with no additional libraries necessary.

3.1 Data

The authors originally used the real-world check-in dataset provided by MIREX[2] (MIREX is a TREC[3]-like evaluation campaign focused on Music IR tasks.) The set includes audio music similarity and retrieval task data which contains 22,074 relevance judgments across 439 queries. It has been suggested by authors to run the greedy selection algorithm to find relevant examples. As a result,

[1] https://github.com/julian-urbano/ecir2015-dcg

[2] http://www.music-ir.org/mirex/wiki/)

[3] https://trec.nist.gov/

they ended up with a total of 4,115 examples that cover 432 unique queries and 5,636 unique documents.

Although not included in the repository the methods described in the original text would be sufficiently clear to retrace the authors steps and create an second version of the user study data.

This process would include creating enough examples to allow for a statistically significant estimations of P (Sat|φ) and a P (Pref |$\Delta\varphi$) based on the user ratings. The original authors chose a to split the satisfaction scores ([0,1]) in to ten increments (bins) and generate 200 examples each. This has to be done for each of the 36 DCG formulations resulting in 7200 examples which would have to be shown at least once to a user.

The authors report a cost of around $0.03 per example on the croud sourcing used by them. Taking into consideration that this cost does not yet take into account unusable user submissions this outgrows what we were willing to pay for a new data set. Urbano Marrero have paid nearly $250 in order to use the service the platform provides and ended up with a data set of only 4115 examples. We were therefore not able to run the experiment on new data.

3.2 Code

The code used to analyze the data and create all plots used in [1] was provided by the authors in a GitHub repository[4]. The repository includes scripts and instructions to run the all files for plot generation and analysis in one command. This however works only if the code is run from a windows machine and the R installation is already at the right location. If another operating system is used, which should be possible due to the underlying R code base being platform independent, directory paths have to be manually changed in each script individually to fit the right installation locations. We therefore chose to run the code directly from the R files in RStudio.

In the original code no algorithms or prediction methods based on randomly splitting the data are used. Therefore the experiment could not be rerun and tested with different random number seeds. However for the same reason all outputs can be considered deterministic, provided that the mathematical calculations are still conducted in the same way as in the original R installation.

The rerunning of the calculations resulted therefore as expected in the exact same results as described in the paper.

4 RESULTS

Regarding user satisfaction it is possible to see a clear pattern across all formulations of discount, satisfaction is underestimated for $\varphi <$ 0.8 and overestimated for bigger scores. Moreover, gain function tend to over emphasize documents with a higher relevance and under emphasize mid-relevant ones, even though users do find useful also the others. With respect to user preference, unless $\Delta\varphi$ > 0.5 users won't pick a clear winner, this shows that P(Pref |$\Delta\varphi$) is proportional to $\Delta\varphi$. Three bias indicators have been computed, b1 indicates how far from P(Sat| φ) = φ the combination is, b2 indicates the goodness of the DCG user mode and b3 how off the combination is from P(Pref |$\Delta\varphi$) = 1. Fig 1 confirms that the more a gain function tends to over emphasize documents the more it will be biased. Furthermore, Linear gain and Constant discount appear to be the least biased.

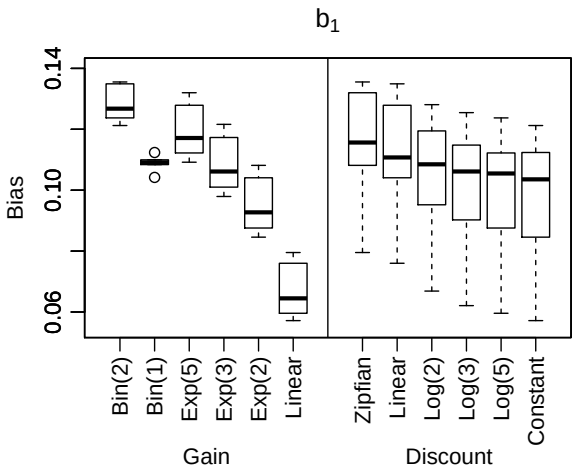

Figure 1: Bias b1

5 CONCLUSION

The results we have reproduced show that exponential gain does underestimate user satisfaction suggest the use of linear gain and constant discount, since as shown in Fig. 2, this combination appears to mitigate the overestimation of satisfaction before φ = 0.8 and its underestimation afterwards.

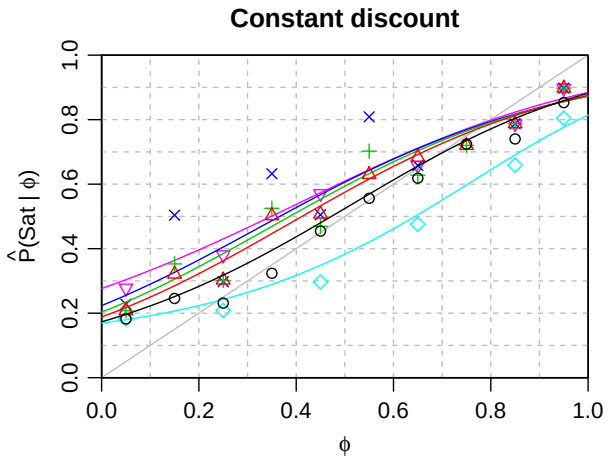

Figure 2: Constant Discount

Future work could be conducted to see whether this result can be generalized also to Text IR tasks. Additionally it would be beneficial to generate new user data by means of another croud-sourcing campaign to reprove the results on different data.

[4]https://github.com/julian-urbano/ecir2015-dcg

REFERENCES

[1] Julián Urbano and Mónica Marrero. 2015. How Do Gain and Discount Functions Affect the Correlation between DCG and User Satisfaction?. In *Advances in Information Retrieval*, Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr (Eds.). Springer International Publishing, Cham, 197–202.