How do Gain and Discount Functions Affect the Correlation between DCG and User Satisfaction?

Julián Urbano[1] and Mónica Marrero[2]

[1] Universitat Pompeu Fabra, Barcelona, Spain
julian.urbano@upf.edu
[2] Barcelona Supercomputing Center, Spain
monica.marrero@bsc.es

Abstract. We present an empirical analysis of the effect that the gain and discount functions have in the correlation between *DCG* and user satisfaction. Through a large user study we estimate the relationship between satisfaction and the effectiveness computed with a test collection. In particular, we estimate 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. We study this relationship for 36 combinations of gain and discount, and find that a linear gain and a constant discount are best correlated with user satisfaction.

1 Introduction

Test collections are used to evaluate how well systems help users in an Information Retrieval task. In conjunction with an effectiveness measure such as Average Precision, they are an abstraction of the search process that allows us to systematically evaluate and improve systems by assessing how good a system is, and which of two systems is better. In particular, collections are an abstraction of the static component in the search process (e.g., documents, topical relevance), while effectiveness measures are an abstraction of the dynamic component (e.g., user behavior, interactions between documents). This user abstraction is advantageous because it makes evaluation experiments inexpensive, easy to run, and easy to reproduce. However, they make several assumptions about how users interact with a system and the perceived utility of the documents it retrieves.

Imagine a system that obtains an effectiveness score $\phi \in [0,1]$ for some query. The best we can interpret ϕ is to assume that $\phi \cdot 100\%$ of users will be satisfied by the system, or $P(Sat|\phi) = \phi$. If we obtain $DCG = 0.85$, we somehow interpret it as 85% probability of user satisfaction. Similarly, if the difference between two systems A and B is $\Delta\phi > 0$, we expect users to agree and prefer A over B. In fact, we expect them to do so *regardless* of how large $\Delta\phi$ is, or $P(Pref|\Delta\phi) = 1$. If the test collection tells us that A is superior to B, we expect users to agree. The extent to which these interpretations are valid depends on whether the assumptions mentioned above hold or not. For instance, relevance judgments are subjective, meaning that we should expect $P(Pref|\Delta\phi) < 1$. Similarly, different effectiveness measures are based on different user models and thus result in different ϕ scores, so $P(Sat|\phi) = \phi$ is not necessarily true.

We present a novel method to investigate these relationships, and study the specific case of DCG in a music recommendation task with informational queries. Through a user study where subjects told us which of two systems they preferred, we empirically map DCG scores onto $P(Sat)$ and $P(Pref)$. An analysis of these mappings for 6 gain and 6 discount functions suggests that the usual exponential gain underestimates satisfaction, and that all forms of discount do so too.

2 Formulations of *DCG*

Let $\mathcal{L}=\{0,1,\ldots,n_{\mathcal{L}}-1\}$ be the set of $n_{\mathcal{L}}$ relevance levels used to make judgments, and let $r_i \in \mathcal{L}$ be the relevance given to document i. The Discounted Cumulative Gain at k documents retrieved is $DCG@k = \sum_{i=1}^{k} g(r_i) \cdot d(i)$, where $g: \mathcal{L} \to \mathbb{R}^{\geq 0}$ is a monotonically increasing *gain* function to map a relevance level onto a utility score, and $d: \mathbb{N}^{>0} \to \mathbb{R}^{>0}$ is a monotonically decreasing *discount* function to reduce utility as documents appear down the ranking. The original formulation used linear gain $g(\ell) = \ell$ and logarithmic discount $d(i) = 1/\max(1, \log_2 i)$ [4]. However, the choice of functions is open. The de facto formulation in IR uses exponential gain $g(\ell)=2^{\ell}-1$ to emphasize the utility of highly relevant documents, and $d(i)=1/\log_2(i+1)$ to penalize all but the first document retrieved [3].

A drawback of DCG is that the upper bound depends on k, $\mathcal{L}$, g and d. $nDCG$ was proposed to normalize scores dividing by the DCG score of an ideal ranking of documents [4]. However, $nDCG$ does not correlate well with user satisfaction when there are less than k highly relevant documents, because systems inevitably retrieve non-relevant documents among the top k [1]. To normalize $DCG@k$ between 0 and 1, we divide by the maximum theoretically possible with k documents. This formulation is better correlated with user satisfaction because it yields $DCG@k = 1$ only when all k documents have the highest relevance:

$$DCG@k = \frac{\sum_{i=1}^{k} g(r_i) \cdot d(i)}{\sum_{i=1}^{k} g(n_{\mathcal{L}}-1) \cdot d(i)}$$

In our experiments we study 6 different gain functions: Linear $g(\ell)=\ell$, Exponential $g(\ell)=b^{\ell}-1$ with bases $b=2$, 3 and 5, and Binary $g(\ell)=I(\ell \geq \ell_{min})$ with minimum relevance $\ell_{min}=1$ and 2. We also study 6 variants of discount: Zipfian $d(i)=1/i$, Linear $d(i)=(k+1-i)/k$, Constant $d(i)=1$ (i.e. null), and Logarithmic $d(i)=1/\log_b(b+i-1)$ with bases $b=2$, 3 and 5. Note that the Constant discount reduces DCG to Precision with Binary gains and to CG with the rest.

3 Methods and Data

We ran an experiment with actual users that allowed us to map system effectiveness onto user satisfaction. Similar to Sanderson et al. [7], subjects were presented with different examples, each containing a query and two ranked lists of results as if retrieved by two systems A and B. Subjects had to select one of

these options: system A provided better results, system B did, they both provided *good* results, or they both returned *bad* results. Behind the scenes, we know the relevance of all documents, so the effectiveness scores ϕ_{A} and ϕ_{B} are known. Subjects indicating that both systems are *good* suggest that they are satisfied with both ranked lists, meaning that ϕ_{A} and ϕ_{B} translate into user satisfaction; if they indicate that both systems are *bad*, it means that they do not translate into satisfaction. Subjects that show preference for one of the systems suggest that there is a difference large enough to be noticed, meaning that $\Delta\phi_{\mathsf{AB}}$ translates into users being more satisfied with one system than with the other. Whether this preference agrees with $\Delta\phi_{\mathsf{AB}}$ depends on which system they prefer.

To compute reliable estimates of $P(Sat|\phi)$ and $P(Pref|\Delta\phi)$ we needed enough examples to cover the full range of ϕ and $|\Delta\phi|$ scores for all 36 DCG formulations under study. To do so, we split the $[0, 1]$ range in 10 equally sized bins, and randomly generate examples until we have at least 200 per bin and DCG formulation. We used an iterative greedy algorithm that at each iteration selects the bin and formulation with the least examples so far, generates a new example for that case, and then updates the corresponding bin in the other formulations.

As search task, we used music recommendation, where the query is the audio of a song and the result of the system is a ranked list of songs deemed as similar (relevant) to the query. This choice has several advantages over a traditional text search task for our purposes. First, it is a purely informational task where the user wants as much relevant information (similar songs) about the query as possible, which makes it a good choice to study $DCG@k$. Second, it is a task known to be enjoyable by assessors and that does not require much time per judgment, considerably reducing assessor fatigue [6]. Third, because subjects have to actively listen to the returned documents, their preferences are not confounded by other factors such as document titles and result snippets. The queries and documents are music clips 30 seconds long, taken from the corpus used in the MIREX audio music similarity and retrieval task (MIREX is a TREC-like evaluation campaign focused on Music IR tasks; see http://www.music-ir.org/mirex/wiki/). We used data from the 2007–2012 editions, comprising 22,074 relevance judgments across 439 queries. After running the greedy selection algorithm, we ended up with a total of 4,115 examples covering 432 unique queries and 5,636 unique documents. As per the task guidelines, all judgments are made on a scale with $n_{\mathcal{L}}=3$ levels, and systems retrieve $k=5$ documents (see [8] for details and the task design).

User preferences for all 4,115 examples were collected via crowdsourcing, as this has been shown to be a reliable method to gather this kind of relevance judgments [6], and it offers a large and diverse pool of subjects to help us generalize results. We used the platform Crowdflower to gather user preferences, as it provides quality control that separates good from bad workers by means of trap examples, as in [7,6,8] (some examples have known answers, provided by us, to estimate worker quality). We manually selected 20 trap questions with answers uniformly distributed. We collected only one answer per example because we are interested precisely in the user variability, not in an aggregated answer reflecting the majority preference. We paid $0.03 per example; the total was nearly $250.

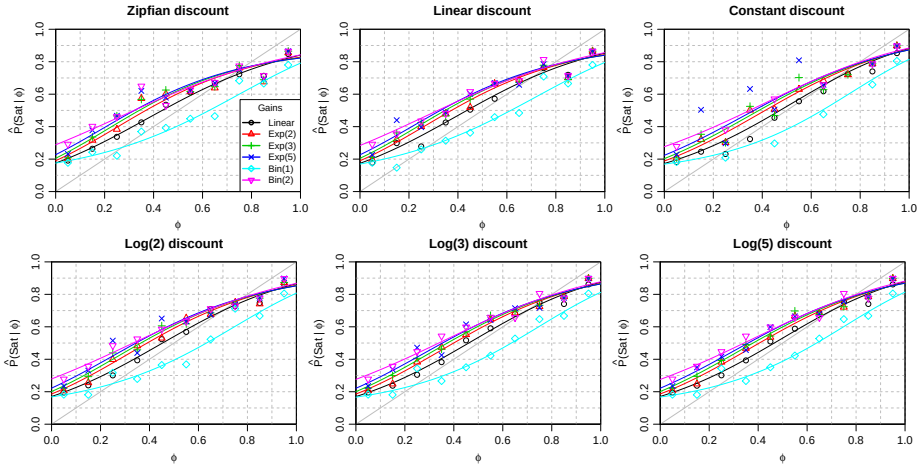

Fig. 1. $\hat{P}(Sat|\phi)$ estimated with 4,050 ranked lists judged as *good* or *bad* and all *DCG* formulations. Points show averages within bins of ϕ, lines show a quadratic logit fit.

4 Results

A total of 547 workers provided 11,042 answers in less than 24 hours. Crowdflower only trusted 175 workers (32%); their trust scores ranged from 73% to 100%, with an average of 90%. After removing answers to trap questions, 113 unique workers were responsible for the answers to our 4,115 examples.

User Satisfaction. For 2,025 of the 4,115 examples (49%) subjects judged both systems as equally good or bad, so we have 4,050 ranked lists judged as satisfactory or unsatisfactory. Fig. 1 shows the estimate $\hat{P}(Sat|\phi)$ for these examples and all *DCG* formulations. The pattern is extremely similar across discount functions: satisfaction is underestimated for low ϕ scores and overestimated beyond $\phi \approx 0.8$. This suggests that users do not discount the utility of documents based on their rank. Within discount functions, we see a subtle but clear pattern as well: gain functions that emphasize highly relevant documents tend to underestimate user satisfaction. For instance, Bin(2) is mostly above the diagonal because only documents with relevance 2 are considered useful by the gain function; those with relevance 1 are deemed as useless, though users did find them useful to some extent. Notice that the exact opposite happens with Bin(1). Similarly, we can see that exponential gains tend to underestimate proportionally to the base. Highly relevant documents are assumed to be much more useful than others (more so with larger bases), so the gain function inherently penalizes mid-relevants because they are not as relevant as they could *supposedly* be.

User Preferences. For 2,090 of the 4,115 examples (51%) subjects indicated that one system provided better results than the other one; whether those preferences agree with the sign of $\Delta\phi_{\mathsf{AB}}$ depends on the *DCG* formulation. Surprisingly, Fig. 2 shows that $P(Pref|\Delta\phi)$ is proportional to $\Delta\phi$, rather than always 1 as we expected. This means that users tend to agree with the test collection,

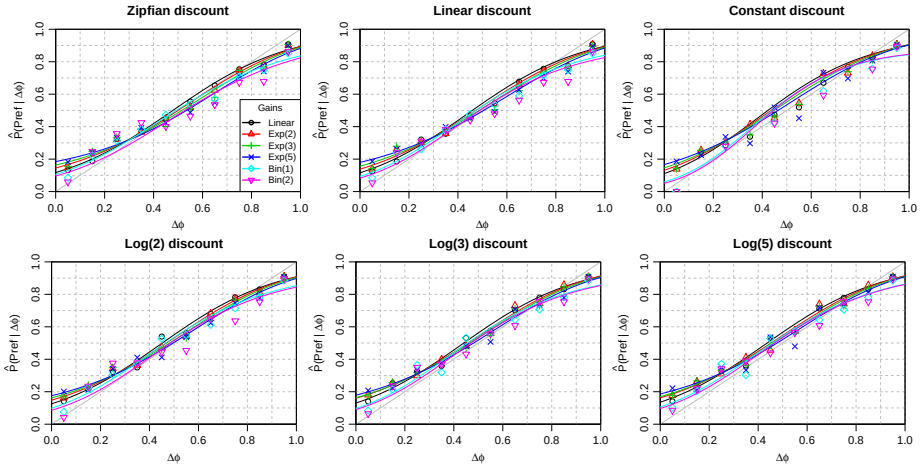

Fig. 2. $\hat{P}(Pref|\Delta\phi)$ estimated with 2,090 examples judged with a preference. Points show averages within bins of $|\Delta\phi|$, lines show a quadratic logit fit.

but differences in effectiveness need to be quite large for the majority of users to do so. On average, unless $\Delta\phi \gtrsim 0.5$ users just can not decide. A subtle but clear pattern appears again: gain functions that overemphasize highly relevant documents work better with low $\Delta\phi$ scores, as the mid-relevant documents that make that difference are found to be more useful than the gain function predicts.

To further analyze what functions correlate best with satisfaction, we computed three bias indicators. The first one, $b_1 = \int |\hat{P}(Sat|\phi) - \phi| \; d\phi$, tells how much off the ideal $P(Sat|\phi) = \phi$ we are in Fig. 1 (note that a large b_1 bias score does not necessarily mean that the DCG formulation is bad; it is just not as easy to interpret as expected). The second one, $b_2 = [\hat{P}(Sat|0) + 1 - \hat{P}(Sat|1)]/2$, tells how large the gaps are at the endpoints $\phi = 0$ and $\phi = 1$ in Fig. 1 (it captures user disagreement and the goodness of the DCG user model). The third indicator, $b_3 = \int 1 - \hat{P}(Pref|\Delta\phi) \; d\Delta\phi$, tells how far apart from the ideal $P(Pref|\Delta\phi) = 1$ we are in Fig. 2 (it measures user discriminative power). For all indicators, an ANOVA analysis shows significant differences among gain and discount functions. Fig. 3 shows that bias is proportional to the emphasis that gain functions give to highly relevant documents, and the steepest discounts are consistently more biased. The Linear gain and Constant discount are the least biased overall.

5 Conclusion

We presented a method to study how well effectiveness measures correlate with user satisfaction, and applied it for a music recommendation task with DCG and a range of gain and discount functions. Our results show that the usual choice of exponential gain underestimates user satisfaction, and that all types of discount tend to do so too, reflecting that users do not pay attention to the ranking. However, the apparent lack of discount effect could be due to the small cutoff used in

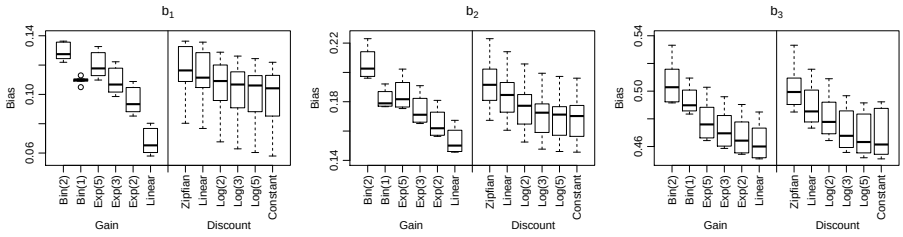

Fig. 3. Bias distributions for all 36 combinations of gain and discount functions.

this task, or the high level of engagement often presented by its users. We also found that differences in DCG need to be large for users to actually agree with the result of a test collection as to which of two systems is better. This suggests that traditional practice of looking at system rankings (e.g. Kendall's τ) and point null hypotheses in statistical significant testing (e.g. $H_0 : \Delta\phi = 0$) oversimplifies the evaluation problem. In qualitative terms, our results largely agree with previous work on both user satisfaction [2,1,7] and reliability of DCG [5].

Future work will investigate the relationship between user satisfaction and system effectiveness for Text IR tasks, as the results presented here do not necessarily generalize. In particular, we will study several other measures, especially for navigational queries and diversity. A similar mapping onto user satisfaction would allow us to evaluate systems within the framework of $P(Sat)$ and $P(Pref)$ for all types of query. Currently we can compute ERR for navigational queries and DCG for informational queries, but averaging all scores together might not be appropriate since they measure effectiveness on different scales. Under a common framework of expected user satisfaction, this problem could be mitigated.

Acknowledgments. Work supported by an A4U postdoctoral grant and the Spanish Government (HAR2011-27540). We thank the reviewers for their comments.

References

1. Al-Maskari, A., Sanderson, M., Clough, P.: The Relationship between IR Effectiveness Measures and User Satisfaction. In: ACM SIGIR (2007)
2. Allan, J., Carterette, B., Lewis, J.: When Will Information Retrieval Be 'Good Enough'? In: ACM SIGIR (2005)
3. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to Rank Using Gradient Descent. In: ICML (2005)
4. Järvelin, K., Kekäläinen, J.: IR Evaluation Methods for Retrieving Highly Relevant Documents. In: ACM SIGIR (2000)
5. Kanoulas, E., Aslam, J.A.: Empirical Justification of the Gain and Discount Function for nDCG. In: ACM CIKM (2009)
6. Lee, J.H.: Crowdsourcing Music Similarity Judgments using Mechanical Turk. In: ISMIR (2010)
7. Sanderson, M., Paramita, M.L., Clough, P., Kanoulas, E.: Do User Preferences and Evaluation Measures Line Up? In: ACM SIGIR (2010)
8. Urbano, J., Downie, J.S., Mcfee, B., Schedl, M.: How Significant is Statistically Significant? The case of Audio Music Similarity and Retrieval. In: ISMIR (2012)