Identifying Complex Patterns in Online Information Retrieval Processes

Debora Di Caprio[1], *Francisco J. Santos Arteaga*[2]

[1] University of Trento, Trento, 38122, Italy

[2] Free University of Bolzano, Bolzano, 39100, Italy

ABSTRACT

We define a computable benchmark framework that replicates the online information retrieval behavior of users as they proceed through the alternatives ranked by a search engine. Through their search processes, decision makers (DMs) must evaluate the characteristics defining the alternatives while aiming to observe a predetermined number satisfying their subjective preferences. A larger number of predetermined alternatives requires more complex information assimilation capacities on the side of DMs. Similarly, the complexity of the algorithms defined to formalize the subsequent retrieval behavior increases in the number of alternatives considered. The set of algorithms delivers two different strings of data, the pages clicked by the DMs and a numerical representation of each evaluation determining their retrieval behavior. We illustrate how, even when providing an Artificial Neural Network with both strings of data, the model faces considerable problems categorizing DMs correctly as their information assimilation capacities are enhanced.

Intelligent Human Systems Integration (IHSI 2022): Integrating People and Intelligent Systems
https://openaccess.cms-conferences.org/#/publications/book/978-1-7923-8988-7

Keywords: Information Retrieval, Satisficing, Click Through Rates, Uncertainty, Heuristics, Artificial Intelligence

INTRODUCTION

Search engines provide a consistent flow of data describing the information retrieval behavior of users as they perform search queries. Eye-tracking technology has validated the fact that users evaluate the alternatives displayed sequentially, according to the ranking delivered by the engines (Lewandowski and Kammerer, 2020). Indeed, the behavior of users displays a consistent trend through time, exhibiting a substantial bias towards the first two alternatives composing the ranking (Dean, 2019; Advanced Web Ranking, 2021). This type of retrieval behavior follows from the trust placed by users on the rankings provided by the search engines (Epstein and Robertson, 2015).

At the same time, the different types of constraints faced by decision makers (DMs) when retrieving and assimilating information have been consistently emphasized in the literature (Di Caprio et al., 2014; Victorelli et al., 2020). That is, the click through rates (CTRs) observed can be generated by retrieval processes involving different degrees of cognitive complexity, ranging from simple independent evaluations to complex processes requiring a completely structured search framework based on a predetermined number of satisfying alternatives.

The current paper defines retrieval algorithms of different complexity to illustrate numerically the fact that aiming to observe two satisfying alternatives within the first ten delivered by a search engine provides a sufficient approximation to the behavior observed. Adding a third alternative allows the algorithm to mimic the CTRs observed, a feature that remains valid as additional alternatives are added to define the retrieval behavior of DMs. Finally, we illustrate how an Artificial Neural Network (ANN) faces substantial difficulties to categorize DMs correctly as the complexity of their retrieval processes is incremented. This is the case even when the ANN is provided with the pages clicked by the DMs and a numerical representation of each evaluation determining their retrieval behavior.

CONTRIBUTION

The complexity inherent to the information retrieval processes of DMs is intuitively illustrated through the evaluation frameworks presented in Figures 1 and 2. Both these figures describe a retrieval scenario where DMs are set to evaluate a total of two alternatives from those composing the initial page of results delivered by an engine. DMs click on those alternatives whose evaluations are above a subjectively predetermined threshold value. For illustrative purposes, the reference thresholds are given by the certainty equivalent values assigned to each alternative, that is, the evaluations that provide DMs with the expected utility derived from a stochastic decision.

Figure 1 describes a standard binary tree where DMs only evaluate the first two alternatives composing the ranking and conclude the retrieval process right after. These alternatives may satisfy the subjective preferences of the DM or may not. However, independently of the results derived from the evaluations, the DM concludes the search after retrieving information from two alternatives. The tree is composed by four final nodes, i.e., vectors describing the alternatives clicked, and three binary decision nodes. Note that, whenever an evaluation is not performed or does not lead to a click, a value of zero has been assigned to the corresponding vector entry.

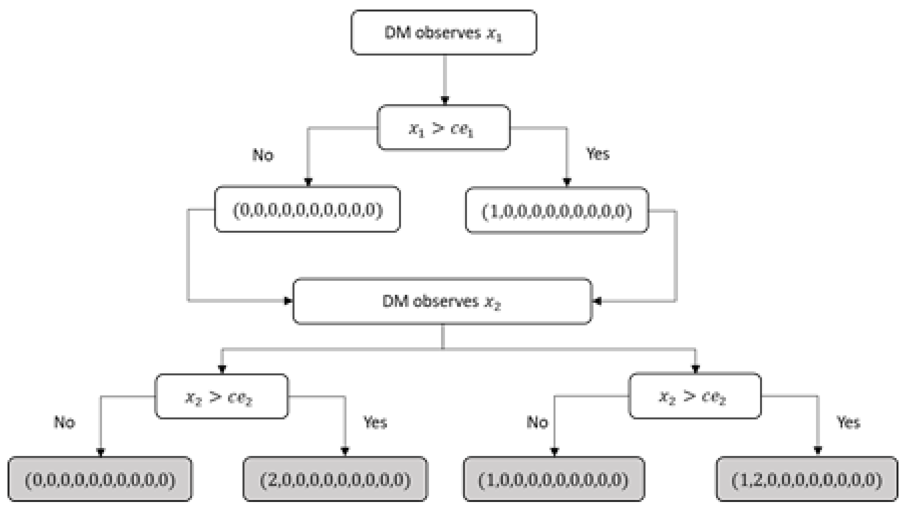

Figure 1. Binary evaluation of the first two alternatives ranked by a search engine

This scenario is quite different from one where DMs aim at observing two satisfying alternatives from a total of ten composing the initial page of search results. In this case, the alternatives selected by the DMs are not necessarily the first two within the ranking. Thus, the information retrieval process requires considering the evaluations that may be observed as DMs proceed through the whole ranking of alternatives. The resulting decision tree is composed by 56 final nodes and 55 binary decision nodes. Clearly, the information assimilation capacities of DMs and the subsequent complexity of the corresponding algorithms differ considerably between both evaluation scenarios.

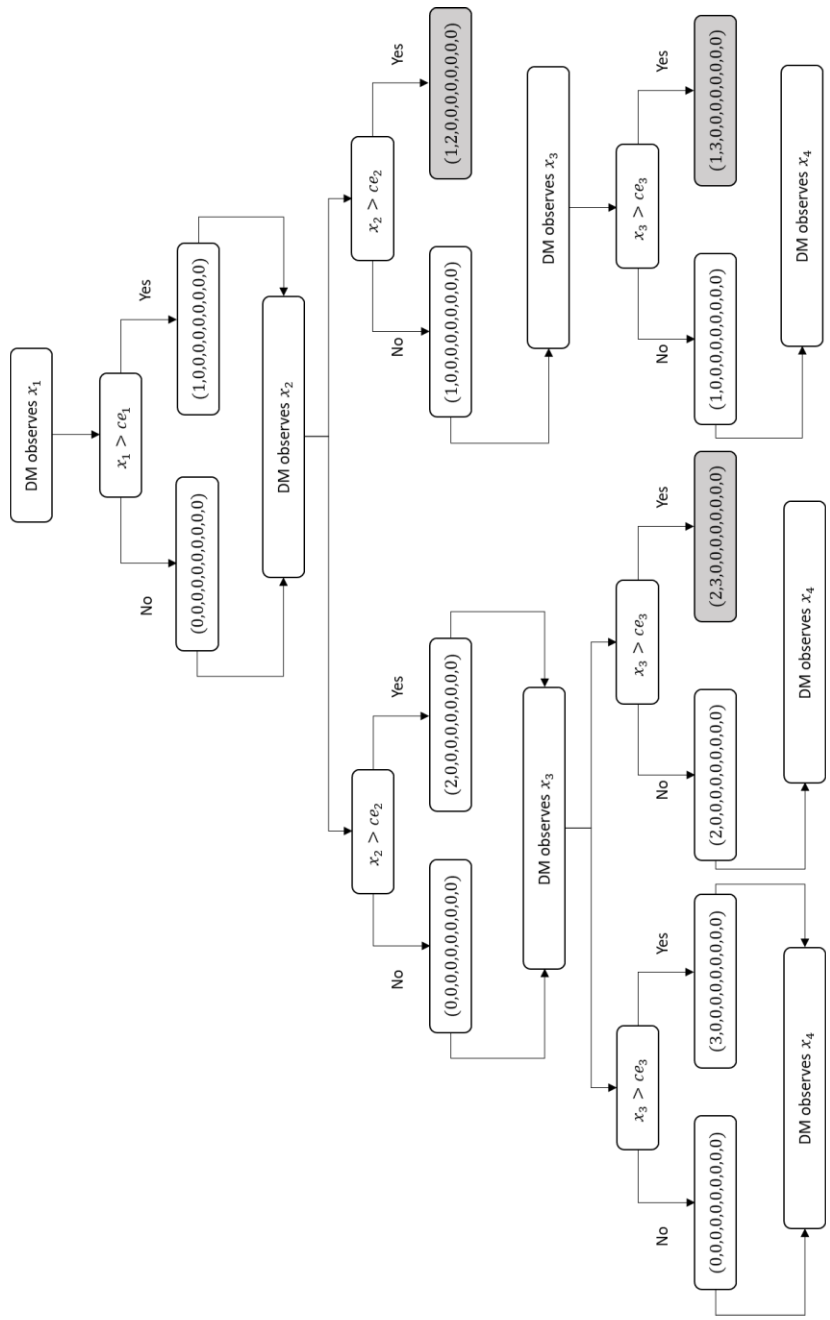

Figure 2. Initial section of a decision tree describing the retrieval behavior of DMs who aim at observing two satisfying alternatives out of a total of ten

Figure 2 presents the initial section of a decision tree illustrating the retrieval process of a DM who aims at observing two satisfying alternatives out of a total of ten. We will illustrate numerically how this retrieval scenario provides a sufficient approximation to the empirical CTRs of users (Dean, 2019; Advanced Web Ranking, 2021). Adding a third alternative delivers an almost identical set of CTRs, a mimicking quality that prevails as alternatives are added up to include the ten ranked within the initial page of search results.

NUMERICAL RESULTS

The CTR of a given alternative is defined as the number of users clicking on the link to the alternative divided by the total number of users performing a search. That is, CTRs reflect the satisfying capacity of the alternatives composing the ranking based on their relative positions. The sum of the corresponding percentages across alternatives does not necessarily add up to one, since users may click on more than one alternative per search query performed.

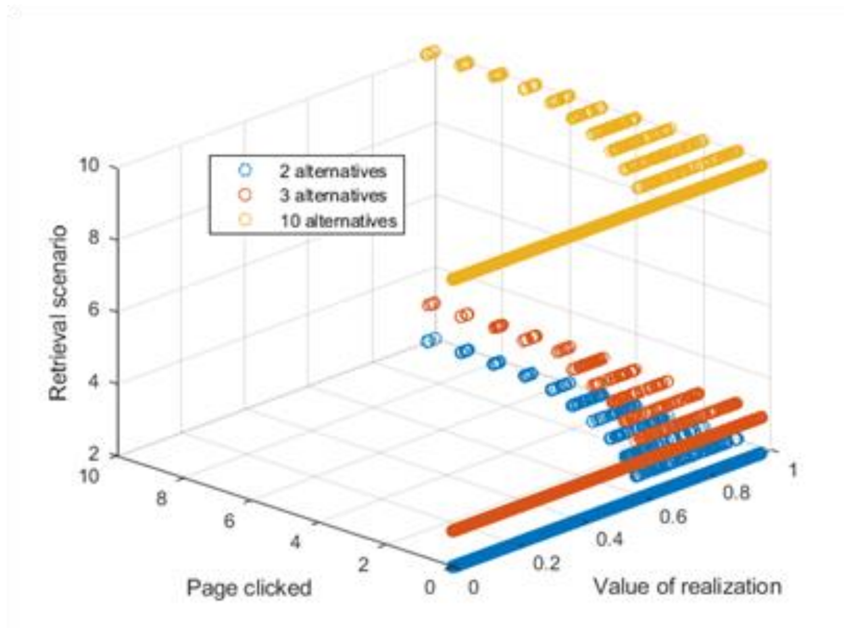

Figure 3. Evaluation patterns generated by increasingly complex information retrieval scenarios

Figure 3 illustrates the evaluation profiles generated by 300 runs of three different retrieval scenarios whose threshold values are determined by the empirical CTRs reported in Dean (2019). Each run amounts for up to 10 clicks, depending on the scenario analyzed and the stochastic realizations of the evaluations, uniformly distributed within [0, 1], relative to the corresponding threshold values. As can be observed, the scenario based on two satisfying alternatives delivers a similar retrieval framework to the limit case with ten alternatives. Thus, we do not require DMs to be endowed with complex information assimilation capacities to generate the CTRs observed empirically.

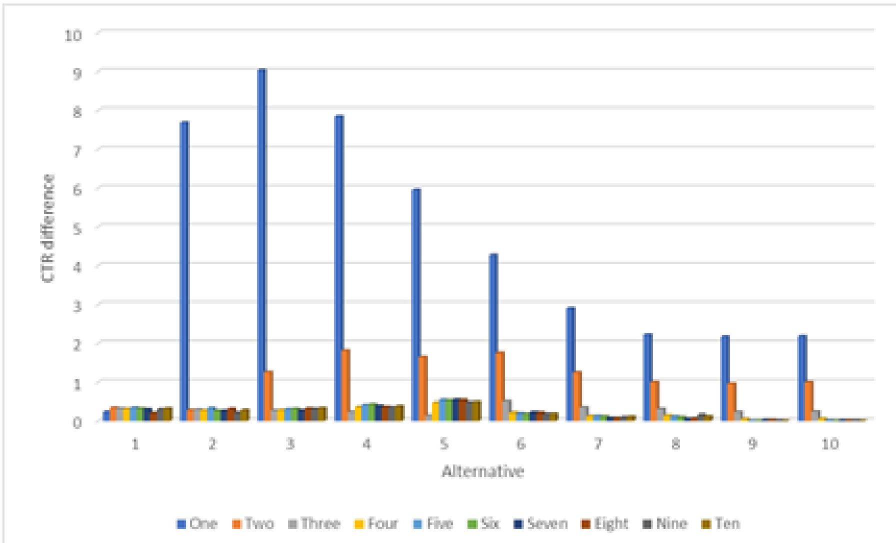

Figure 4. Absolute value of the difference between the CTRs reported by Dean (2019) and those obtained from one million runs of the current algorithms per information retrieval scenario

This intuition is validated in Figure 4, which describes the absolute value of the difference between the CTRs reported by Dean (2019) and those obtained from one million runs of the current algorithms per retrieval scenario. Note how the scenario based on one satisfying alternative fails to provide a sufficiently close approximation to the empirical CTRs. More importantly, the difference in CTRs remains below two percentage points throughout the scenario with two alternatives. The differences between the CTRs generated in the remaining scenarios and those described by Dean (2019) remain marginal, with the algorithms providing almost identical retrieval profiles in all cases.

Given the similarity of the profiles obtained when considering three alternatives or more, up to a total of ten, we investigate whether an ANN endowed with the retrieval profiles of these users can categorize them correctly. The profiles include both the pages clicked by the DMs and a numerical representation of each evaluation determining their retrieval behavior.

Figure 5 presents the results obtained and highlights the difficulties faced by the ANN when categorizing 100,000 DMs from each retrieval scenario (70% of the data has been used for training, 15% for validation, and 15% for testing). These difficulties become particularly evident when the information assimilation capacities of DMs are enhanced so as to allow them to consider five satisfying alternatives or more within their evaluation processes.

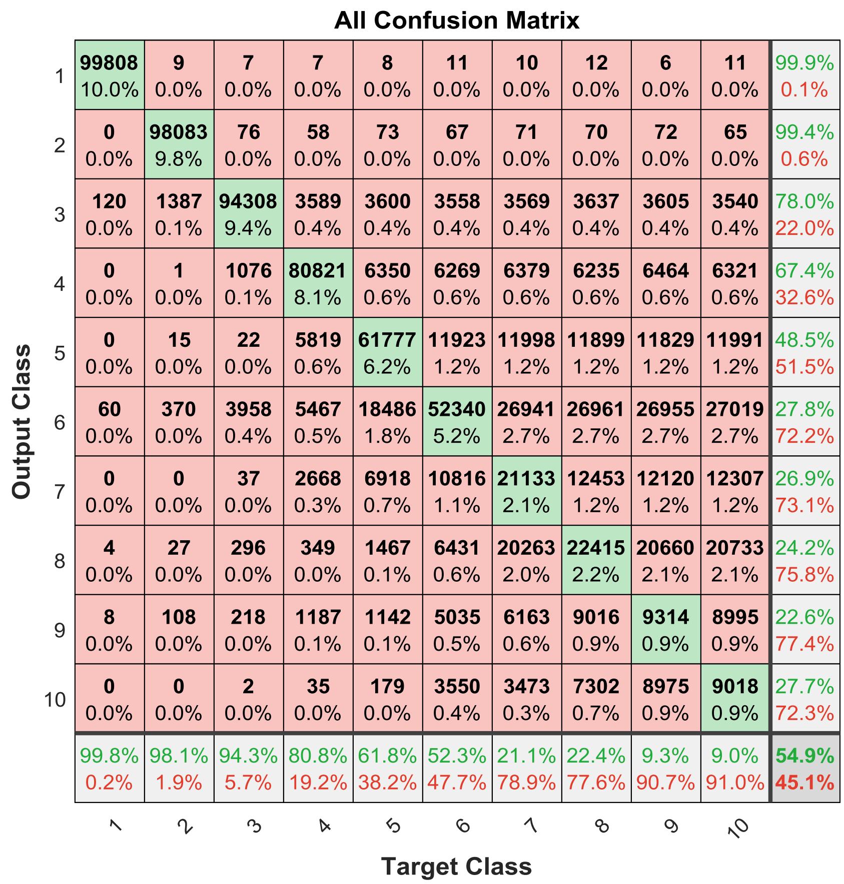

Figure 5. ANN confusion matrix categorizing 100,000 DMs per information retrieval scenario

CONCLUSIONS

We have defined a set of retrieval algorithms of different complexity and illustrated how they mimic the behavior of DMs who aim at observing at least three alternatives satisfying their subjective preferences among those ranked within the initial page of results provided by a search engine. This result has been extended to illustrate the difficulties faced by an ANN to categorize DMs according to their information retrieval capacities when the latter aim at observing five or more satisfying alternatives within the ten composing the first page of results delivered by the engine.

REFERENCES

Advanced Web Ranking: Google organic CTR history. Last accessed September 23rd, 2021. (2021) https://www.advancedwebranking.com/ctrstudy/

Dean, B.: We Analyzed 5 Million Google Search Results. Here's What We Learned About Organic Click Through Rate (2019) https://backlinko.com/google-ctr-stats

Di Caprio, D., Santos-Arteaga, F.J., Tavana, M.: The Optimal Sequential Information Acquisition Structure: A Rational Utility-Maximizing Perspective. Appl. Math. Model. 38, 3419--3435 (2014)

Epstein, R., Robertson, R.E.: The Search Engine Manipulation Effect (SEME) and Its Possible Impact on the Outcomes of Elections. PNAS 112, E4512--E4521 (2015)

Lewandowski, D., Kammerer, Y.: Factors Influencing Viewing Behaviour on Search Engine Results Pages: A Review of Eye-Tracking Research. Behav. Inf. Technol. 40, 1485--1515 (2020)

Victorelli, E.Z., Dos Reis, J.C., Hornung, H., Prado, A.B.: Understanding Human-Data Interaction: Literature Review and Recommendations for Design. Int. J. Hum. Comput. Stud. 134, 13--32 (2020)