An Empirical Study on Clarifying Question-Based Systems

Zou, J.; Kanoulas, E.; Liu, Y.

DOI
10.1145/3340531.3412094

Publication date
2020

Document Version
Final published version

Published in
CIKM ’20

License
Article 25fa Dutch Copyright Act

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)
An Empirical Study on Clarifying Question-Based Systems


1 INTRODUCTION

One of the key components of conversational search and recommender systems [5, 7] is the construction and selection of good clarifying questions to gather item information from users in a searchable repository. Most current studies either collect and learn from human-to-human conversations [1, 2, 4], or create a pool of questions on the basis of some "anchor" text (e.g., item aspects [5], entities [6–8], grounding text [3]) that characterizes the searchable items themselves. Although the aforementioned works have demonstrated success in helping systems better understand users, most of them evaluate algorithms by the means of simulations which assume users are willing to provide answers to as many questions as the system generates, and that users can always answer the questions correctly, i.e. they always know what the target item should look like in its finest details. On the basis of such assumptions, their evaluations (e.g., Zhang et al. [5], Zou et al. [6], Zou and Kanoulas [7]) focus on whether the system can place the target item at a high ranking position. To the best of our knowledge, there is no empirical validation of whether and to what extent users can respond to these questions, and the usefulness perceived by users while interacting with the system.

In this paper we conduct a user study by deploying an online question-based system to answer the following research questions: (1) To what extent are users willing to engage with a question-based system? (2) To what extent can users provide correct answers to the generated questions? (3) How useful do users perceive while interacting with a question-based system? We believe that answering these research questions can help the community design better evaluation frameworks and more robust question-based systems.

2 STUDY DESIGN

In our study, the users interact with a question-based system in the domain of online retail. The user is answering questions prompted by the system with a “Yes”, a “No” or a “Not Sure”, in order to find a target product to buy. The architecture of our system is shown in Figure 1, with the user going through 4 steps.

Step 1: Category selection. In this step, the users select an Amazon category 1 that they feel most familiar with to fit their interests, e.g. a category from which they have purchased products before.

Step 2: Target product assignment. We randomly assign a target product to the user from the selected category. The user is requested to read the title and the description of the product carefully. A picture of the product is also provided. This simulates a use case in which the user really knows what she is looking for, as opposed to an exploratory use case. If the user is not familiar with the target product the user can request a new product.

Step 3: Find the target product. After the user indicates that the conversation with the system can start, the target product disappears from the screen and the system selects a question to ask to the user. The user needs to provide an answer on the basis of the target product information she read in the previous step. Once the user answers the question, a 4-by-4 grid of the pictures of the top sixteen ranked products is shown to the user, along with the next clarifying question. The user can stop answering questions at any time when

---

1 Categories and dataset we used: http://jmcauley.ucsd.edu/data/amazon/links.html
she wants to stop during her interaction with the system. To select what clarifying question to ask, a state-of-the-art algorithm [7] is deployed to first extract important entities from each product description (e.g., product aspects) and construct questions in the form of “Is [entity] relevant to the product you are looking for?”. Then, it selects to ask the information-theoretically optimal question, that is the question that best splits the probability mass of predicted user preferences over items closest to two halves, and updates this predicted preference on the basis of the user’s answer [7]. In this work, we update the predicted preference using the correct answer, i.e. the answer which agrees with the description of the product, independent of the user’s answer. In other words, we study the user behavior under a perfect system from an information theoretical point of view, leading to a best-case analysis and conclusions.  

Step 4: Questionnaire. In this step users are asked a number of questions about their experience with the system for further analysis.

3 EXPERIMENTS AND ANALYSIS

3.1 Research Questions

Our research questions revolve around the user engagement and perceived value of the system:  

Q1 Are users willing to answer the clarifying questions, how many of them, when do they stop and why, and how fast do they provide the answers?  

Q2 To what extent can users provide correct answers given a target product, and what factors affect this?  

Q3 How useful do users find the clarifying questions, what is their overall experience, and how likely is it to use such a system in the future?

3.2 Participants

Prior to the actual study, we ran a pilot study with a small number of users, in a controlled environment, and iterated over the experimental design, and the user interface until no issues or concerns were reported. For the actual study 53 participants located in the USA were recruited through Amazon Mechanical Turk and 1025 conversations were collected. The participants were of varying gender, age, career field, English skills and online shopping experience. In particular, gender: 34 male, 19 female; age: 2 in 18-23, 8 in 23-27, 14 in 27-35, 29 older than 35 years old; career field: 22 in science, computer & technology, 8 in management, business & finance, 7 in hospitality, tourism, & the service industry, 3 in education and social services, 2 in arts and communications, 2 in trades and transportation, 9 did not specify their career field; English skills: all of them were native speakers; online shopping experience: 44 were mostly shopping online, 9 did online shopping once or twice per year. Participants were paid 2.5 dollars to complete the study. Also, we only engaged Master Workers ¹, filtered out those users who spent less than 3 seconds on reading the product title and descriptions, and users who gave random answers (~50% correct/wrong), for quality control.

3.3 User Willingness to Answer Questions

In RQ1, we first investigate whether users are willing to answer the system’s questions and how many of them, both by observing the actual number of questions users answered when interacting with the question-based system and what they declared at the exit questionnaire. The results in Figure 2 show that users answer a minimum of 2 and a maximum of 48 questions. The average number of questions answered per target product is 11.4, the median is 7, and 70.3% of users answered 4-12 questions per product, while at the exit questionnaire about 50% of the users declare that they are willing to answer 6-10 questions. Further, we explore why users stop answering questions. Users could select one out of six answers during the exit questionnaire: “The target product was

¹High performing workers identified by Mechanical Turk who have demonstrated excellence across a wide range of tasks.
We conclude that the percentage of incorrect answers varies across different question characteristics, or one could develop question selection strategies that also account for the time spent on each question. The results show that users usually spend more time when they are not sure about the answers, but almost the same time when they are right or wrong about a question.

### 3.5 User Perceived Helpfulness

Regarding RQ3, we explore how useful do users perceive while interacting with such a question-based system. We ask the user (a) whether they think the question-based system is helpful, (b) whether they will use such a system in the future, and (c) what their rating is for the system, ranging from 1 (very negative) to 5 (very positive). The results using oracle answers are shown in Figure 5, in the three plots respectively. From the results we collected, most users think the question-based system is helpful and they will use it in the future. Specifically, 83.9% of users are positive about using such a system in the future, 30.4% of users are neutral, and 8.9% of users are negative. Regarding user ratings, the results show 46.5% of 5-star ratings, 37.5% of 4-star ratings, 7.1% of 3-star ratings, 7.1% of 2-star ratings, and 1.8% of 1-star ratings. 84% of the users gave a rating at least as high as a 4.

### 4 CONCLUSION AND DISCUSSION

In this paper we conduct an empirical study using a question-based product search system to gain insight into user behavior and interaction with such systems. We deploy a state-of-the-art question-based system online and collect interactive log data and questionnaire data for analysis. We find that users are willing to answer a certain number of the system generated questions and stop answering questions when they find the target product, only if the questions are relevant and well-selected. While users are able to answer these questions effectively, they also provide incorrect answers at a rate...
of about 17%, which varies across users, products and question characteristics. Last, most users are positive towards question-based systems, and think that these systems help them towards achieving their goals. The take-home message, if there is one, is that current research should drop the assumption that users are happy to answer many questions as the system generates and that all questions are answered correctly.

One limitation of this work is the isolated clarifying-based environment of the study. A more realistic experiment would require clarifying questions to be embedded in an existing environment, where the user is enabled to not only answer questions, but also reformulate her query or filter results by selecting pre-defined item attributes, and browse the results to the preferred depth. Also a mixed-initiative approach under which a system switches from asking questions, to understanding user searches, and combining the two is worth studying. A further limitation of this work is the fact that this was not an in-situ experiment but a simulation of a use case of a question-based system by involving crowd workers. Hence, the findings are as good as our simulation of a user looking for a target product. Other factors, such as question quality, question format, and noisy answers, may affect the results, and studying therefore these factors in an A/B testing experiment would be beneficial. We leave all these as future work.

Acknowledgements

This research was supported by the NWO Innovational Research Incentives Scheme Vidi (016.193.501), the NWO Smart Culture - Big Data / Digital Humanities (314-99-301), the H2020-EU.3.4. - SOCIETAL CHALLENGES - Smart, Green And Integrated Transport (814961) and the China Scholarship Council program.

REFERENCES