Crowdsourcing Backstories for Complex Task-Based Search

Manuel Steiner
RMIT University
Melbourne, Australia
manuel.steiner@student.rmit.edu.au

Damiano Spina
RMIT University
Melbourne, Australia
damiano.spina@rmit.edu.au

Falk Scholer
RMIT University
Melbourne, Australia
falk.scholer@rmit.edu.au

Lawrence Cavedon
RMIT University
Melbourne, Australia
lawrence.cavedon@rmit.edu.au

ABSTRACT

Backstories provide vital contextual information for information retrieval evaluation. They are useful as textual representations of information needs, for example to aid in relevance judgements as part of test collections for performance evaluation, for studying longer search queries, and for interactive retrieval. While backstories exist for some popular search tasks and domains thanks to evaluation campaigns such as TREC, NTCIR, and CLEF, they are not available for a large range of other tasks and domains. In this paper, we explore crowdsourcing as an approach for obtaining high-quality backstories, with the aim of supporting the development of backstories as key resources for new domains and search tasks. Compared to typical crowdsourcing tasks in the IR domain, such as gathering relevance judgements or short textual search queries, obtaining backstories is more complex. Workers are required to think of information need scenarios and put these thoughts into comparatively lengthy text fragments. This possibly entails a higher cognitive load and longer working time. We describe a crowdsourcing methodology to maximise the usefulness of results, using the creation of backstories for the job search domain as an example. We also present and release a collection of 756 job search backstories, which was obtained via the proposed methodology.

CCS CONCEPTS

• **Information systems** → **Query intent**; **Specialized information retrieval**; *Test collections.*

KEYWORDS

information need, backstory, crowdsourcing

ACM Reference Format:
Manuel Steiner, Damiano Spina, Falk Scholer, and Lawrence Cavedon. 2021. Crowdsourcing Backstories for Complex Task-Based Search. In *Australasian Document Computing Symposium (ADCS '21), December 9, 2021, Virtual Event, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3503516.3503526

ADCS '21, December 9, 2021, Virtual Event, Australia

ACM ISBN 978-1-4503-9599-1/21/12…$15.00
https://doi.org/10.1145/3503516.3503526

1 INTRODUCTION

Search processes are generally driven by information needs [26], which capture the current environment and situation of the user. Based on the information need, a user formulates queries (for example, using keywords) that are then submitted to an information retrieval system, which returns a list of documents.

While users querying an information retrieval system can judge result relevance based on their current information need, an explicit representation is useful in other cases, such as performance evaluation. The traditional method to evaluate search engine performance is to employ test collections consisting of queries, result documents and relevance judgements, which are compared to results returned by the tested system [25]. Well established test collections, such as the ones used in TREC [28], CLEF [11], and NTCIR [14], are available for certain popular retrieval tasks. However, for other domains or types of tasks, such as job search, few or no resources are available. In order to establish collections for domain-specific search, both search topics (information need statements) and relevance judgements need to be created. To judge the relevance of results for complex information needs, backstories, can constitute a vital contextual information – as opposed to search queries only, which may not convey complete needs. In a typical job search process, a user issuing the keyword *manager* would possibly find most returned job advertisements about manager positions relevant, based on the search query only. However, the user's current context influences their search intent and adds complexity to the search. Factors, such as location or work experience can restrict the set of documents the user perceives as relevant. As backstories describe information needs in textual form [6], they allow for an explicit representation of search intents and potentially aid in tasks, such as query disambiguation, personalising search or judging document relevance in evaluation settings. Test collections could benefit from built-in ambiguity [24]. A variability in backstories to express multiple information needs for a single query would aid in creating such resources.

Since generating backstories by conducting user studies or via human experts is time-consuming and costly (based on wages), we explore crowdsourcing as a potential cost-effective alternative. Additionally, this method might produce a higher variety of backstories. Crowdsourcing has extensively been used for successfully gathering relevance judgements [2–4, 16] or collecting search queries [5]. To our knowledge, the use of crowdsourcing to collect backstories that represent complex information needs in written forms has not yet been explored. In this context, crowd workers are required

to think of information needs, and then formulate these as text. This is usually substantially longer than other crowdsourced text, such as search queries. Although creating backstories is arguably more complex than some more typical crowdsourcing tasks, the approach should be feasible with well controlled settings.

The contributions of this paper are two-fold. First, we describe a crowdsourcing methodology to obtain written information needs of high quality in an efficient, cost-effective manner. Gathered backstories can then be used, for example to obtain relevance judgements for offline effectiveness evaluation. Second, we instantiate the methodology to create a collection of backstories describing job search information needs. The collection consists of 756 backstories associated with 50 job search queries, which we make publicly available to the research community. To the best of our knowledge, there exists no comprehensive collection of backstories for the job search domain. We present the creation of backstories for a job search scenario as a motivating example. The crowdsourcing methodology however is potentially suitable for other complex search settings as well.

The remainder of the paper is organised as follows. Section 2 discusses related work. Our proposed methodology to obtain backstories for job search scenarios via crowdsourcing is detailed in Section 3. Section 4 presents and discusses the results of applying the proposed methodology to create a collection of crowdsourced backstories for job search. We conclude with a summary and future work in Section 5.

2 RELATED WORK

This section outlines existing literature about backstories, personas, and crowdsourcing in information retrieval settings.

Backstories in Information Retrieval. Backstories – i.e., descriptions of information needs in textual forms – are an integral part in an alternative evaluation method proposed by Borlund and Ingwersen [6]. In comparison to TREC, where topic narratives provide instructions about what is considered relevant [28], Borlund and Ingwersen let users formulate their own search queries and determine relevant results in order to simulate a realistic search process. Follow-up research demonstrated that real information needs can be substituted with simulated ones [7]. Bailey et al. [5] manually created backstories for TREC topics. They then crowdsourced multiple queries for each backstory, which led to a test collection with query variability. As part of an experiment to evaluate the relationship between search query length and information need specificity, Phan et al. [21] had participants think of backstories and suitable search queries.

Personas as Alternative. A commonly used technique to model users as well as their information needs and system interactions in human computer interaction is *Personas* [9, 13, 22]. Personas are elaborate constructs that represent fictive people and their context (such as family information and hobbies). Backstories are best compared to Persona scenarios, which describe how goals are accomplished by interacting with the system in question. Some information of Personas might be useful for information need statements, depending on the search domain. However, especially for evaluation purposes, a broad range of different information needs are required. Establishing these with backstories in a concise text snippet is potentially more efficient and cost-effective compared to creating a large number of Personas and scenarios.

Crowdsourcing in Information Retrieval. Various crowdsourcing aspects have been researched extensively in information retrieval settings [1, 8, 10, 12, 18, 19, 23]. Alonso et al. [4] outlined how to leverage Amazon Mechanical Turk to obtain relevance judgements for result documents. Restricting workers based on their qualifications and getting multiple judgements for a single document yields higher quality results. Similarly, Alonso and Baeza-Yates [2] documented effective methods for using crowdsourcing to gather relevance judgements, including keeping tasks as short as possible and implementing thorough instructions. Quality and accuracy of crowdsourced relevance judgements in comparison to expert assessors has been investigated, noting valuable insight into justification of judgements [3, 16]. Changes in quality of relevance judgement crowdsourcing work by varying pay, worker qualification and required worker effort was investigated by Le et al. [17]. Kazai [15] explored quality impacts of using data with known answers (gold data) throughout tasks. Vliegendhart et al. [27] concluded that a well-designed task title helps to convey task expectations to workers in tasks exhibiting high imaginative load. Furthermore, free-text justification questions, tailored to the task in question, aid in successful crowdsourcing of information requiring a certain degree of imagination. Compared to the work presented in this document, existing crowdsourcing literature overwhelmingly focuses on obtaining relevance judgements or short search queries. Such tasks potentially require less cognitive attention from workers.

3 GATHERING BACKSTORIES VIA CROWDSOURCING

We now explain an iteratively developed crowdsourcing methodology to gather backstories for pairs of queries and results, using the job search domain as an example. The last iteration represents the final settings that resulted in maximisation of useful backstories. The interface of the final task is depicted in Figure 1.

3.1 Crowdsourcing Setup

For obtaining backstories that describe search scenarios for which provided queries (e.g. extracted from search logs) are a suitable representation, the platform Figure Eight[1] (formerly CrowdFlower) was utilised. Workers were presented with a search query in conjunction with a target job advertisement. The data was originally mined from a query log of a major job search engine, covering the period from May 2017 to June 2017.[2] The query log is not publicly available due to the data being proprietary. The job advertisement associated with each query was randomly sampled from the documents that were retrieved at least once in the top 100 results for the given search query. This makes the job advertisements potentially but not necessarily relevant to the search queries. The data set has been used in a different experiment to obtain relevance judgements via crowdsourcing. In this work, we used the provided relevance

[1] https://www.figure-eight.com/

[2] No personally identifiable information was made available or used, and this investigation was performed under Ethics Application at RMIT University.

judgements for quality control (as detailed in the section Third Iteration). Presented information from the job advertisements contained classifications (location, work type, field of employment and salary), as well as the advertisement text. The title was omitted in order to avoid creation of backstories based on the title alone without consideration of full job advertisements. Workers were asked to carry out the following steps.

- They had to carefully read the provided information and think of an information need which a searcher could have in mind when submitting the given query, and receiving the job advertisement as a result.
- They were then requested to put that backstory in writing (at least 100 characters, to avoid blank submissions or extremely short texts) into a text area with the label *"Write a backstory that would make the job advertisement satisfactory to the given search request."*.
- They had to indicate how well they thought their generated backstory matches the query and job advertisement on a Likert scale with ratings from one to five. The radio button group was labelled *"How satisfactory is the job advertisement in relation to the backstory that you wrote?"*.
- To provide a highly individualised justification question with free-form text answer, as suggested by Vliegendhart et al. [27], McDonnell et al. [20] and Kutlu et al. [16], workers were asked to justify why they chose the rating of how well the backstory suits the provided information (minimum 100 characters). A text area with the label *"Please explain your satisfaction decision."* was provided for this input.

3.2 Iterative Task Refinement

The task was refined in three iterations. In each one, 15 queries and job advertisements were used to collect a total of 75 backstories (5 per query and job advertisement). As mentioned, gathering multiple backstories for a single search query allows for the creation of test collections that contain different information needs for a single query. The provided information was randomly sampled from the previously described query log. To minimise potential language barriers, the task was only made available to workers from Australia, Canada, Ireland, New Zealand, United Kingdom and United States of America. One contributor was able to work on at most ten tasks, and the minimum working time on a single task was set to three minutes to ensure that workers read the provided information carefully. The task was titled *"Create Scenarios (Backstories) For Job Seekers"*.

We define a usable backstory as English text that contains an information need from the perspective of a job seeker. The information need has to be suitable for the query and result provided in the task. The text however can include minor grammatical, spelling or punctuation mistakes.

First Iteration. No additional accuracy checks or worker filtering were employed beyond those described previously (only workers from countries with English as native language, work on at most 10 tasks, minimum time of 3 minutes). Payment was initially set to US$0.50 per task. This is higher than the usual few cents for relevance judgements [4] due to the complexity and the longer working time this task possibly entails. A single page (work unit) consisted of a single task. The worker confidence level was set to three out of three; this setting defines how confident Figure Eight is in the accuracy of workers, based on their work history; the higher the level, the smaller the pool of people, and the higher the historical accuracy.

The results of this iteration led to around 70% of the generated backstories being of acceptable quality. To determine usability, one author manually analysed the stories, i.e. read them and checked for English sentence structure and conveyance of a job search information need relevant to the provided information. Unacceptable backstories could be categorised as follows. Around 17% were written from the hirer's perspective, rather than the job-seeker's perspective, and thus are not suitable for the specific search task being studied. Another 12% consisted of phrases that were copied and pasted from the provided job advertisements, and 1% were meaningless text (resembling sentences but not information needs).

Second Iteration. The instructions were revised based on the analysis of generated backstories from the first iteration. They were rewritten to more clearly indicate that backstories should reflect the perspective of the job seeker. This was also included in the label for the backstory input field with the phrase *"Imagine the situation of a job seeker!"*. To further clarify what was required, a hint (placeholder text) was added to each text input, in order to remind workers of the type of content they should provide (*"What is the current background situation of the job seeker?"*) for the backstory field and *"Why did you choose the above rating for your backstory?"* for the justification field).

To address the copy and paste issue, the ability to use copy from and paste into text input fields was disabled. Further, the usage of substrings solely from the job advertisement as text in the input fields was detected via a JavaScript method, and such substrings were not accepted as input. Payment was increased to US$0.65 based on worker feedback.

Surprisingly, the proportion of usable backstories decreased to around 63%. However, almost all of the unsuitable backstories (35%) consisted of text phrases that were repeated across responses as a sort of template response (a combination of both job searcher and hirer perspective) with only minor variations to adapt to the provided information. A total of eleven accounts from a single worker channel submitted this sort of phrase, which could indicate a single person was using multiple accounts, or that people were exchanging text fragments. Such texts were classified as meaningless, because they resembled English sentences, but not information needs. A further 1% were incomprehensible text.

Third Iteration. To mitigate the reuse of template-style text phrases by workers – a problem identified in the second iteration – a similarity check of generated backstories was introduced. To enable this checking, the number of tasks on a page was increased to two. The backstories that were entered for both tasks on a single page were compared to each other; if they were highly similar (Levenshtein similarity greater than 0.65), the input was not accepted.

Analysis of the first two iterations also revealed that workers wrote rather short backstories that closely followed the provided examples in terms of content; the length was often close to the minimum requirement. In order to encourage more elaborate backstories, the minimum number of characters per story was increased

to 200. The example stories provided as part of the instructions for workers were also rewritten to contain considerably longer texts compared to the minimum requirements as well as additional information, such as searcher history or personality. In previous iterations they only contained the immediate information need.

An additional check was introduced by employing a "quiz mode", which presents an initial entry task to filter workers before they are permitted to earn money, and enables regular checks on worker accuracy. Such filtering processes are a well studied technique to increase the usability of crowdsourcing results [4, 17]. For relevance judgements, the information that is crowdsourced can be used with predefined answers to check worker accuracy. This approach however is impractical for obtaining backstories because there is no unique text fragment for a combination of search query and result. Based on the information need, a variety of backstories can be written. The task thus needs to be extended with a suitable question, which might not be related to the actual data being gathered.

To detect workers that randomly provide unusable information, they were required to additionally indicate if the provided job advertisement fits the query via a binary response (with the label *"Is the job advertisement relevant to the search query?"*). One of the two tasks on a page contained a known relevance answer to this question. The response could be compared to the relevance judgements available in the data set used for our task; for this purpose, we used 20 query and job advertisement pairs where the job advertisement was either clearly relevant or irrelevant for the query and appropriate justification was provided as to why the result was relevant or not. This check should filter workers with the intent of randomly providing unusable information.

Due to the increased complexity of the task, and to therefore enable a larger pool of eligible workers to be eligible overall, the workforce confidence level in the crowdsourcing platform was set to two out of three.

For this third iteration, the proportion of usable backstories increased to 91%. The remaining 9% were backstories written from the hirer's perspective.

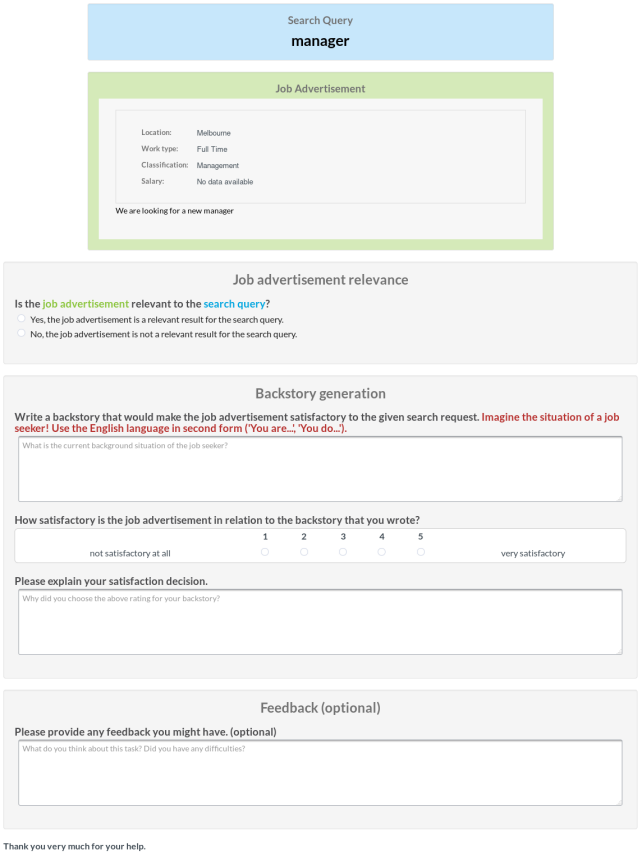

Figure 1: The final interface of the crowdsourcing task. Note that a page contains two tasks.

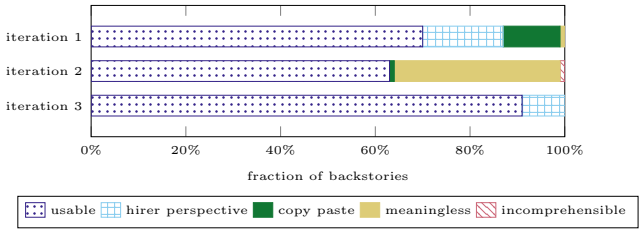

Figure 2: Fractions of backstories per type and iteration.

4 RESULTS AND DISCUSSION

We now provide descriptive measures for the iterations of our crowdsourcing setup and analyse obtained backstories. Furthermore, we discuss the final collection of backstories for job search scenarios, which was obtained via the proposed methodology.

4.1 Qualitative Analysis

Figure 2 summarises the distribution of backstories for each crowdsourcing task iteration described above, split into the categories: (i) usable, (ii) hirer's perspective (a hirer's information need), (iii) copy paste (text copied from the job advertisement), (iv) meaningless (not resembling an information need), and (iv) incomprehensible.

With the additional worker filtering and similarity checks in the third iteration, the percentage of usable backstories improved compared to the first two, despite activating a larger pool of workers with lower confidence. Interestingly, this entailed a substantially longer total run-time (6h, 4h, and 85h for each iteration, respectively), possibly due to the fact that 29 out of 68 workers failed the initial filtering task and thus were not permitted to work on further tasks. The majority of users generated a low number of backstories (median 2 for each iteration), probably because the task is more time-consuming compared to other work. The average times to complete tasks were 4m55s, 5m4s and 5m27s, which suggest workers were not hampered by the checks added in iterations two and three. Keeping the increase of minimum backstory length for the third iteration in mind, the generated backstories were generally close to the minimum required length (median of 30, 32 and 50 words), even though we provided quite elaborate examples in the third iteration to motivate the writing of longer phrases.

Descriptive statistics of the three iterations are depicted in Table 1. An increase in usable backstories can be observed with a

Table 1: Descriptive statistics of the three iterations. The cost of iteration three include tasks that were used for accuracy checks. The backstory lengths are based on usable backstories.

Iter.	*Total Cost (US$)*	*Usable Stories (%)*	*Overall Completion Time (h)*	*Mean Task Completion Time*	*# Workers*	*Stories per Worker*			*Story Length (words)*		
						Min	*Median*	*Max*	*Min*	*Median*	*Max*
1	45	70	6	$4m55s$	21	1	2	10	16	30	62
2	68	61	4	$5m04s$	31	1	2	10	16	32	76
3	171	91	85	$5m27s$	68	1	2	10	33	50	149

Table 2: Example backstories that were obtained via the third iteration.

Query	*Crowdsourced Backstory*
mining sales	You have 10 years experience in the sales sector including 5 years as management. You are very effective at spurring growth and have been working in the mining sector for the past two of those years. Due to relocation you would like to find a sales position in the Perth area.
welder	You are a certified welder, familiar with every tool and are looking for work in a factory environment. You have over 7 years experience working in a factory and are a capable machine operator as well.
barista	You are looking for a part time position as a barista after having worked in such a role in a previous location prior to moving to a new town for attending university. You have had experience in supporting the manager, working as team leader of a small team of baristas and waiting staff.

trade-off being a higher overall completion time, possibly due to the amount of workers only completing a low number of tasks, for the third iteration.

Examples of crowdsourced backstories are presented in Table 2. They generally resemble job search scenarios and are suitably detailed for use in further tasks, such as search scenarios and making corresponding relevance judgements. In contrast to the creativity of some workers, others tend to follow examples quite closely and thus their resulting backstories have a similar form. Despite providing different examples, we suspect that many workers might only read the first one and follow the structure of the presented backstory. In order to have consistently formatted backstories, a final normalisation step might be necessary to account for minor grammatical and spelling issues.

In addition to manually reviewing backstories, we experimented with the usage of readability scores as potential measures to automatically discard unusable backstories. However, initial investigations showed no apparent trends that such metrics would be suitable for automatic classification.

4.2 Job Search Backstory Collection

Using the same setup as in the third iteration (as described in the previous section), we obtained backstories for 50 search queries and associated target job advertisements. The same data source as for developing the methodology was used to extract search queries and job advertisements (i.e. query logs of a major job search platform during the period of May 2017 to June 2017).

In comparison however, the job advertisement was selected as follows. The job advertisement must (i) be retrieved as part of the result set in response to the submitted query, (ii) be present in the top 20 results (the first result page), (iii) be clicked at least once after the query was submitted. The latter two restrictions should enhance the potential of the selected job advertisement being relevant to submitted search query. A click signal on a job advertisement on the first result page is potentially a feasible indicator of a searcher's interest in the document.

The search queries were extracted based on an informed methodology to gather different query types (ones that are submitted to the search engine based on potentially a single intent and ones that can be used for possibly multiple intents). The extraction process is based on query reformulation frequency as a potential alternative to click entropy for detecting possibly ambiguous queries. Queries were gathered in this way to allow for further research (analysing backstories that were generated for different query classes). The exact method is omitted in this document because it is not pertinent to the presented crowdsourcing methodology and the collection of backstories in general.

A total of 756 backstories were collected for the query and job advertisement pairs. They were obtained during the period from 6 September 2018 to 3 February 2019. 71 workers took part in the crowdsourcing task. The rather long time-frame it took to obtain the collection is possibly due to the same reasons the third iteration took much longer to finish than the previous ones during the methodology development phase. There seems to be a trade-off between speedier but poor quality and slower but higher quality results due to added restrictions imposed on crowdsourcing workers. Table 3 shows descriptive statistics of the obtained backstory collection. The resulting backstories exhibit similar properties to obtained backstories from the third iteration, with slightly longer task completion times.

The collection of backstories and associated search queries is publicly available for research purposes. The resources can be accessed via https://github.com/manuelsteiner/job-search-backstories. Backstories were manually reviewed and corrections were applied to achieve a collection of high quality. Text snippets not suitable as an information need statement were removed. Grammatical and spelling errors were rectified and missing punctuation was inserted. Furthermore, in order to normalise the way, backstories are written, the text snippets were changed to second form in the instances

Table 3: Descriptive statistics of the job search backstory collection.

Number of queries	50
Number of backstories	756
Minimum number of backstories per query	10
Average number of backstories per query	15
Maximum number of backstories per query	28
Minimum number of backstories per worker	2
Median number of backstories per worker	4
Maximum number of backstories per worker	40
Minimum task completion time	1m45s
Average task completion time	6m40s
Maximum task completion time	15m0s
Minimum backstory length (words)	8
Median backstory length	45
Maximum backstory length	234

where different or mixed forms were used by crowdsourcing workers.

5 SUMMARY AND FUTURE WORK

We propose a crowdsourcing methodology to effectively gather backstories to support the creation of resources for the study of complex and domain-specific search tasks such as job search. In such tasks, backstories are essential as they provide a more comprehensive representation of the information need than short text representations such as queries. To our knowledge, this work provides the first description for crowdsourcing backstories for complex search tasks. A collection of 756 backstories for job search – created with the proposed methodology – is publicly available for research purposes at https://github.com/manuelsteiner/job-search-backstories.

Future work includes the use of the established backstory collection to obtain relevance judgements for job search evaluation. We also plan to investigate methods for automated assessment of backstory usefulness and quality (e.g., for query disambiguation purposes). Various instructions for crowdsourcing workers, such as the request to use second form, can be investigated for better effort estimations. Finally, we believe our proposed crowdsourcing methodology can be applied to create resources to better understand the role of search in other complex task scenarios such as real-estate, automobile, or other e-commerce domains.

ACKNOWLEDGMENTS

This research was partially supported by Australian Research Council Project LP150100252 and SEEK Ltd.

REFERENCES

[1] Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2019. ComQA: A Community-sourced Dataset for Complex Factoid Question Answering with Paraphrase Clusters. In *Proc. NAACL-HLT*. 307–317.
[2] Omar Alonso and Ricardo A. Baeza-Yates. 2011. Design and Implementation of Relevance Assessments Using Crowdsourcing. In *Proc. ECIR*. 153–164.
[3] Omar Alonso and Stefano Mizzaro. 2012. Using crowdsourcing for TREC relevance assessment. *Inf. Proc. & Man.* 48, 6 (2012), 1053–1066.
[4] Omar Alonso, Daniel E. Rose, and Benjamin Stewart. 2008. Crowdsourcing for relevance evaluation. *SIGIR Forum* 42, 2 (2008), 9–15.
[5] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2016. UQV100: A Test Collection with Query Variability. In *Proc. SIGIR*. 725–728.
[6] Pia Borlund and Peter Ingwersen. 1997. The development of a method for the evaluation of interactive information retrieval systems. *J. Documentation* 53, 3 (1997), 225–250.
[7] Pia Borlund and Peter Ingwersen. 1999. The Application of Work Tasks in Connection with the Evaluation of Interactive Information Retrieval Systems: Empirical Results. In *Proc. MIRA: Evaluating Interactive Information Retrieval.*
[8] Paul D. Clough, Mark Sanderson, Jiayu Tang, Tim Gollins, and Amy Warner. 2013. Examining the Limits of Crowdsourcing for Relevance Assessment. *IEEE Internet Computing* 17, 4 (2013), 32–38.
[9] Alan Cooper and Paul Saffo. 1999. *The Inmates Are Running the Asylum.* Macmillan Publishing Co.
[10] Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In *Proc. WSDM*. 162–170.
[11] Nicola Ferro. 2019. What Happened in CLEF…For a While?. In *Proc. CLEF*. 3–45.
[12] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2021. The Impact of Task Abandonment in Crowdsourcing. *IEEE Trans. Know. Data Eng.* 33, 5 (2021), 2266–2279.
[13] Bernard Jansen, Joni Salminen, Soon-gyo Jung, and Kathleen Guan. 2021. Data-driven personas. *Synthesis Lectures on Human-Centered Informatics* 14, 1 (2021), i–317.
[14] Noriko Kando, Kazuko Kuriyama, and Toshihiko Nozue. 1999. NACSIS Test Collection Workshop (NTCIR-1) (poster abstract). In *Proc. SIGIR*. 299–300.
[15] Gabriella Kazai. 2011. In Search of Quality in Crowdsourcing for Search Engine Evaluation. In *Proc. ECIR*. 165–176.
[16] Mucahid Kutlu, Tyler McDonnell, Yassmine Barkallah, Tamer Elsayed, and Matthew Lease. 2018. Crowd vs. Expert: What Can Relevance Judgment Rationales Teach Us About Assessor Disagreement?. In *Proc. SIGIR*. 805–814.
[17] John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. 2010. Ensuring quality in crowdsourced search relevance evaluation. In *Proc. SIGIR*. 17–20.
[18] Joel M. Mackenzie, Rodger Benham, Matthias Petri, Johanne R. Trippas, J. Shane Culpepper, and Alistair Moffat. 2020. CC-News-En: A Large English News Corpus. In *Proc. CIKM*. 3077–3084.
[19] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. 2017. On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation. *ACM Trans. Inf. Sys.* 35, 3 (2017), 19:1–19:32.
[20] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. In *Proc. HCOMP*. 139–148.
[21] Nina Phan, Peter Bailey, and Ross Wilkinson. 2007. Understanding the relationship of information need specificity to search query length. In *Proc. SIGIR*. 709–710.
[22] John Pruitt and Jonathan Grudin. 2003. Personas: Practice and Theory. In *Proc. Designing for User Experiences (DUX)*. 1–15.
[23] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In *Proc. SIGIR*. 439–448.
[24] Mark Sanderson. 2008. Ambiguous queries: test collections need more sense. In *Proc. SIGIR*. 499–506.
[25] Mark Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Found. & Trends in IR* 4, 4 (2010), 247–375.
[26] C. J. van Rijsbergen. 1979. *Information Retrieval.* Butterworth.
[27] Raynor Vliegendhart, Martha Larson, Christoph Kofler, Carsten Eickhoff, and Johan Pouwelse. 2011. Investigating Factors Influencing Crowdsourcing Tasks with High Imaginative Load. In *Proc. Workshop on Crowdsourcing for Search and Data Mining (CSDM)*. 27–30.
[28] Ellen M. Vorhees and Donna K Harman. 2005. *TREC: Experiment and evaluation in information retrieval.* Vol. 1. The MIT Press.