Identifying Re-finding Difficulty from User Query Logs

Sargol Sadeghi
RMIT University
Melbourne, Australia
seyedeh.sadeghi@rmit.edu.au

Roi Blanco
Yahoo! Research Center
Barcelona, Spain
roi@yahoo-inc.com

Peter Mika
Yahoo! Research Center
Barcelona, Spain
pmika@yahoo-inc.com

Mark Sanderson
RMIT University
Melbourne, Australia
mark.sanderson@rmit.edu.au

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

David Vallet
Universidad AutÃşnoma
Madrid, Spain
david.vallet@uam.es

ABSTRACT

This paper presents a first study of how consistently human assessors are able to identify, from query logs, when searchers are facing difficulties re-finding documents. Using 12 assessors, we investigate the effect of two variables on assessor agreement: the assessment guideline detail, and assessor experience. The results indicate statistically significant better agreement when using detailed guidelines. An upper agreement of 78.9% was achieved, which is comparable to the levels of agreement in other information retrieval contexts. The effects of two contextual factors, representative of system performance and user effort, were studied. Significant differences between agreement levels were found for both factors, suggesting that contextual factors may play an important role in obtaining higher agreement levels. The findings contribute to a better understanding of how to generate ground truth data both in the re-finding and other labeling contexts, and have further implications for building automatic re-finding difficulty prediction models.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Performance, Measurement, Experimentation, Human Factors.

Keywords

Assessor Agreement, Re-finding, Difficulty Detection.

1. INTRODUCTION

Re-finding, where a user is looking for a previously seen piece of information or document, comprises around 40% of web searches [8]. Research has shown that when users have difficulty re-finding, this can be reflected in the way that users formulate their query and navigate search results. So common are these searches that studying re-finding behavior is a key component of improving the user search experience.

Query log analysis has been used to establish indications of user difficulties in search (e.g. see Liu et al. [7]). Previous analyses mostly relied on users' self-assessed perception of task difficulty to evaluate inferred difficulty indications from query logs. However, it is not always practical or feasible to collect such self-reported data. Therefore, it is worth to examining if human assessors can judge difficulties from logs.

There is some evidence from past research, which indicates the feasibility of detecting task difficulty reliably from query logs: (a) The search behavior of users could change by difficulty of the tasks [1]; (b) The main reasons for task difficulties are more reflected in the search process rather than a user's self-assessed factors such as their familiarity with a topic [3]; (c) Although task difficulty is a subjective concept and dependent on the users, Liu et al. [6] reported that there are common factors in the perception of task difficulty among users; (d) The perception of users for task difficulty could change before and after search tasks, as Liu et al. compared against search behavior, which could indicate that what can be inferred from user performance might be more robust indication for task difficulty than what the users perceive.

As a way of exploring how consistently assessors are able to detect user's difficulty, the agreements between assignments can be examined. So far, the only study of agreement in the re-finding from logs has been on the identification of re-finding in email [2]. Human assessments have been studied in many other information retrieval (IR) contexts, such as relevance assessments [10, 11, 12]. Two research questions are addressed: (1) Can human assessors consistently agree on levels of user difficulty in re-finding tasks? (2) Do contextual types of factors affect the level of agreement between assessors in detecting re-finding difficulties?

The main contributions of this work are: (1) to establish the ability of human assessors to agree when labeling re-finding difficulty; (2) examining agreements in terms of assessors and guidelines, and also investigating search behavioral factors. The results contribute knowledge on how to generate ground truth data both in the re-finding and other labeling contexts, and have further implications for building automatic prediction models.

2. EXPERIMENTAL METHODOLOGY

Here, we describe the data sets and their pre-processing.

2.1 Dataset

A sample was taken from 30 days of interactions with the Yahoo! search engine, from the $1^{st} - 30^{th}$ of October 2012. Logged data from 2,847,028 unique anonymized users consisted of the submit-

ADCS '14, November 27–28 2014, Melbourne, VIC, Australia

ACM 978-1-4503-3000-8/14/11 ...$15.00
http://dx.doi.org/10.1145/2682862.2682867.

Table 1: Frequency of exact-last-click pairs in each of the 12 types of paired goals. R_R and R_O are the rank of last clicks in the *potential re-finding* and *original* goals respectively; Potential re-finding goal $< hasmorequeries, hasmoreclicks >, T = True, F = False$.

Effort level / Rank level	1: $< T, T >$	2: $< T, F >$	3: $< F, T >$	4: $< F, F >$
1: $R_R > R_O$	4.13%	0.86%	8.53%	3.56%
2: $R_R = R_O$	21.05%	6.40%	29.40%	13.07%
3: $R_R < R_O$	3.78%	0.99%	3.86%	4.38%

ted queries, the URL, the rank position of clicked search results, and a timestamp for each event. We followed the terms of service and privacy policies of Yahoo!.

In terms of log data segmentation, we focused on *goals*, which are defined as a group of related queries and corresponding clicks submitted by a user to perform a task with an atomic search need [5]. Previous work has shown that goals are more accurate than session timeouts for identifying task boundaries. A goal is identified by predicting boundaries using classifiers based on features indicative of relatedness in sequences of queries (e.g. number of words in common). We used the same approach for identifying goals in this study.

2.1.1 Potential Re-finding Tasks

In past studies, re-finding has typically been defined based on repetitions of clicks on the same URLs across the searches of the same user [8, 9]. Here, we add a constraint that the repeated URLs must be the last clicks of a user's goal (*exact last click*). Past research has shown that the last click in a task is important to capture a relevance signal [13]. Adding this constraint increases the likelihood of including re-finding tasks, the focus of this study.

All goals from the same user were extracted using Jones and Klinkner's technique [5]. Goals were ordered by their timestamp and all possible sequential goals were paired. Pairs that occurred less than thirty minutes apart were not included. Note that this time constraint was not applied for goal identification, and the segmentations therefore included tasks that were possibly interleaved. In total 39,683,301 paired search goals were identified, applying the exact last click constraint resulted in a final set of 2,959,327 pairs. We refer to the first goal in a pair as the *original*, and the second as the *potential re-finding* goal.

2.2 Sampling Data

As the focus of our work was on detecting difficulties in re-finding, we developed a set of filters to remove excessively easy re-finding, such as searching for popular home pages (e.g. Facebook). [1] This left 9,445 exact-last-click paired goals, which were sampled to give a manageable number to assess.

Two key factors that could affect an assessor's perception of difficulty are system performance and searcher behavior. These characteristics were therefore considered when sampling from the pool of exact last click paired goals. For the system factor, the *rank* of each exact last click was noted, along with the sign of the *difference* in the ranks of the last clicks in the re-finding and original goals. To represent searcher behavior, the relative number of queries and clicks between the paired goals was considered and categorized into four classes, as shown in the heading row of Table 1. We randomly sampled ten paired goals from each cell of the table, which were labeled by assessors.

[1] `http://tinyurl.com/navigational-rules`

Number of days between goals: 1
Original Goal (Time: 2012/10/12 19:55)
Q: bleacher report college football T: 2
C(3): `www.cbssports.com/collegefootball` T: 15
C(10): `bleacherreport.com/college-football` *exact last click*
Potential re-finding Goal (Time: 2012/10/13 20:51)
Q: college fottball T: 2
Q: college football T: 9
C(1): `espn.go.com/college-football/` T: 16
C(39): `www.cbssports.com/collegefootball` T: 20
C(43): `bleacherreport.com/college-football` *exact last click*

Figure 1: Example paired goal: Q is query; C(n) is click at rank n; T is dwell time (seconds).

2.3 Labeling Design

When labeling, assessors were presented with extracts from search engine query logs that included: queries, clicked URLs (including their rank), the time between queries and clicks, and the dates on which the two paired search goals were conducted. An example is shown in Figure 1. Assessors were asked two questions: (1) *"Do you think that in the second search the user is re-finding document(s) that were found in the first search?"* (Possible responses were "yes", "no", "not sure".) (2) *"In terms of search difficulty, would you say the second search is?"* (Possible responses were "easy", "difficult", "not sure".)

Re-finding was defined as repeat searching for a document that was previously found. The notion of the *difficulty* was defined for assessors in a broad sense of whether it seems that the user is struggling to find the target document. Specifically assessors were instructed to consider the effort of the user in a) providing input information for searching; b) finding the relevant documents, and c) recognizing the target document.

2.4 Assessors and Guidelines

Twelve assessors were recruited from RMIT university and given *initial guidelines* composed of a set of paired goals (separate from the experimental set) that were selected and pre-annotated by the first author. We investigated whether assessor ratings were affected when they were provided with additional *detailed* guidelines. Based on Webber et al.'s joint-assessed approach [12], two assessors labeled a sample data set together, discussing and proposing more detailed instructions and examples. The examples were organized under different categories including user performance (in terms of query and click indications), system performance (rank information), and temporal information (time between goals), and compiled into a final set of *detailed* guidelines. [2] An example of a detailed guideline was "the time gap between paired goals could affect the level of difficulties".

We also analyzed the effect of an assessor's experience on their ratings. Note that in this study, the notion of experience is defined in terms of the familiarity of the assessors with the labeling task, that is whether they previously conducted the same labeling job (*experienced*), or not (*inexperienced*). There were six experienced assessors, who had conducted an initial labeling exercise on a separate dataset using the initial guidelines.

Four experimental settings were investigated: (1) inexperienced assessors using initial guidelines; (2) experienced assessors using initial guidelines; (3) inexperienced assessors using detailed guidelines; and (4) experienced assessors using detailed guidelines. In each setting, a data set of 120 paired goals were randomly ordered and labeled by each assessor, with three assessors per experimen-

[2] `http://tinyurl.com/detailed-guideline`

Table 2: Mean pair-wise assessor agreement for re-finding identification and difficulty assessment problems. Matching symbols indicate a significant difference ($p < 0.05$) between a pair of settings.

Settings: guidelines, assessors	Re-finding identification	Re-finding difficulty
initial, inexperienced	82.9% ♣♥♠	51.2%♥♠
initial, experienced	94.3% ♣†	61.0%♦†
detailed, inexperienced	99.0% ♥	59.3%♥♦
detailed, experienced	100% ♠†	78.9%♠†

tal setting, which is in line with the number of assessors in related studies (e.g. a study by Webber et al. [12]). Note that all settings used the same data, and experienced assessors were initially trained on a separate sample of 120 pairs. For practical reasons, the assessment process was divided into three blocks of 40 paired goals; each block took about half an hour to complete, and assessors were able to take short breaks between blocks.

3. EXPERIMENTAL RESULTS

In this section we investigate the overall agreement between assessors, and examine effects from varying the guidelines, and assessor experience. We also investigate the effect of perceived search performance and user effort on assessor agreement.

3.1 Overall Agreement

Table 2 shows the agreement for each setting. Using initial guidelines and inexperienced assessors, the mean pair-wise agreement for identifying re-finding was 82.9%, for assessing difficulty agreement was 51.2%. This setting was considered as a baseline. The results from other settings showed greater overall agreement, both from using more detailed guidelines and from more assessor experience.

We also calculate Cohen's kappa (κ) for inter-assessor agreement; the average of pairwise κ scores across all settings was 0.5 for identifying re-finding and 0.2 for the detection of difficulty. Although the level of agreement for difficulty is low in comparison to re-finding, it is still fair considering the levels of assessor agreement in other IR contexts such as relevance judgments [12].

The significance of agreement rates between the settings were also analyzed using McNemar's chi-squared test. We employed McNemar's test instead of an ANOVA test as response values were in a binary scale. Comparing use of detailed guidelines vs. initial ones shows significant differences ($p < 0.05$) for both identification and difficulty assessment. However, in comparing experienced vs. inexperienced settings, agreement rates were significantly different only for the identification of re-finding.

Most agreement rates were found to be significantly different from each other. However, it appears that judgments are not always affected by the experience of assessors. Though, providing more detailed guidelines led to significantly higher agreement rates. This is in contrast with labeling efforts in the TREC Legal Track relevance judging, where use of more detailed guidelines could not significantly increase the level of agreement in comparison to general guidelines [12]. The amount of effort that should be invested by researchers into the development of guidelines at the appropriate level of detail therefore appears to be dependent on the labeling problem that is being considered.

As the identification of re-finding was designed as an obvious easy labeling job, it is considered as an upper bound agreement to which difficulty assignments can be compared. As can be seen by improving guidelines and experience of assessors we could reach closer to the upper bound agreements. Note that in reporting agreements we removed cases where assessors were not able to make a judgement (i.e. "not sure" labels). On average only 3.3% and 1.1% of the rates were labeled as "not sure" for identifying re-finding and detecting difficulty respectively. We plan to investigate the characteristics of these ambiguous cases in future work.

Table 3: Mean pair-wise percentage agreement of re-finding identification and difficulty detection for a) system performance (rank); and b) searcher behavior (effort) factors. The factor levels are from Table 1. Matching symbols indicate a significant difference ($p < 0.05$) between a pair.

(a) rank

Rank level	Re-finding identification	Re-finding difficulty
$R_R > R_O$	93.0%♣	58.3%♥
$R_R = R_O$	95.3%♣	63.1%♦
$R_R < R_O$	93.3%	65.6%♥♦

(b) effort

Effort level	Re-finding identification	Re-finding difficulty
$< T, T >$	90.4%	57.4%
$< T, F >$	95.4% ♦	66.7%
$< F, T >$	92.7% ♦†	62.5%
$< F, F >$	98.0% †	63.3%

3.2 Rank, Effort and Agreement

A second aim of our experiments was to investigate whether different levels of system performance and searcher behavior had an effect on assessor agreement rates. We therefore examined the effect of the rank and effort level features on assessor agreement.

Starting with Table 3(a), for re-finding agreement identification, results suggest that if the target document (last click) in the second search appears at a similar rank compared with the click in the original search, this provides a clue to assessors when assessing re-finding, leading to higher agreement. Otherwise, if the target document is at a lower or higher rank, then the assessor interpretation is more ambiguous, leading to lower agreement. For the re-finding difficulty agreement, it appears that when the rank of the target document does not change or it is higher in the re-finding task, assessors agreed more.

Examining Table 3(b), when considering the number of query and click actions that were carried out in a potential re-finding goal, agreement on re-finding was highest for $< F, F >$ and lowest for $< T, T >$ and $< F, T >$. The latter settings represent re-finding cases with a greater number of clicks compared to the original goal. This could indicate that having greater numbers of clicks makes it hard for assessors to identify re-finding. Consider a case where there were more clicks with high dwell time (potential relevant clicks) in the potential re-finding goal. This makes it hard for assessors to agree if users are searching for specific documents.

The effort level feature appears to affect agreement only on re-finding; there is no significant difference between agreement rates for difficulty assessment. It appears that it is hard to agree on difficulty when users are submitting different numbers of queries and clicks, relative to their original access. A longer search (particularly in terms of click actions) makes agreement more difficult. Other factors such as dwell time could also impact on agreement rates. In comparing agreement on re-finding with difficulty, difficulty is hard for each factor level. Developing more examples in the guidelines should be considered for judges.

4. DISCUSSIONS

Analysis in Section 3 showed there is a fair level of agreement

when assessing difficulty, which is influenced by the level of detail of guidelines along with other contextual search conditions (e.g. rank and effort levels). Investigation of other features, such as temporal characteristics, suggested they may also play a role. These need to be further investigated, particularly for detecting task difficulties, as the importance of search behavioral features has been highlighted in previous work [6]. This is not only a matter of re-finding and difficulty agreements, but also this would be a requirement for other human labeling jobs to identify effective factors along with the selections of guidelines and assessors.

In the context of relevance assessment, assessor disagreement was mainly studied in terms of assessors, guidelines, documents, ranks and topical variance [11]. It would be worthwhile to explore the effect of additional behavioral search factors on the level of agreement between assessors for re-finding, as these factors have been found to be important elsewhere [4]. These underlying factors should be considered in developing guidelines and sampling data, which could lead to higher agreement rates. This could result in more balanced ground-truth data in terms of incorporating multiple factors reflective of the whole dataset. The ground truth data can be used for building predictive models using machine learning techniques, which could be used by search engines for adapting search results by predicting the type of user task.

As the main aim of this study was to measure the level of agreement for detecting re-finding difficulties, the experimental data were sampled from likely re-finding tasks. However, a more general data set can be explored, where paired goals are not necessarily ended with exact last clicks. As another avenue for illustrating assessor's ability in detecting difficulties, the gathered labels from assessors can be examined against a ground truth data generated by searchers. Moreover, gathering qualitative data from assessors after labeling could provide better insight of influential factors, which we are going to explore in future work.

5. CONCLUSIONS

This paper asked the following two research questions: (1) Can human assessors consistently agree on levels of user difficulty in re-finding tasks? (2) Do contextual types of factors affect the level of agreement between assessors? An experiment was conducted where twelve assessors each labeled 120 instances of potential re-finding tasks and difficulties from search logs. A maximum agreement of 78.9% was obtained between assessors when rating re-finding task difficulty. Providing more detailed guidelines was found to significantly improve assessor agreement rates on re-finding and difficulty rates. This is in contrast to previous work on labeling tasks in other contexts, such as the TREC Legal Track relevance assessments, where detailed guidelines did not significantly affect recorded agreement rates.

Two search characteristics representative of system performance and searcher effort were also examined. The analysis indicated significant differences between some levels of examined factors, which can provide a better understanding of human perception of re-finding and task difficulty. These factors can be further explored not only in the context of re-finding, but also for other labeling applications, such as relevance judgments. This knowledge could result in higher agreement rates between human assessors, and consequently more balanced ground truth data to be generalized and extrapolated over the whole domain of the labeling problem.

6. ACKNOWLEDGMENTS

This work was supported in part by Yahoo! and the Australian Research Council's Discovery Projects Scheme (project DP140102655).

7. REFERENCES

[1] A. Aula, R. M. Khan, and Z. Guan. How does search behavior change as search becomes more difficult? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 35–44. ACM, 2010.

[2] D. Elsweiler, M. Harvey, and M. Hacker. Understanding re-finding behavior in naturalistic email interaction logs. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 35–44. ACM, 2011.

[3] J. Gwizdka and I. Spence. What can searching behavior tell us about the difficulty of information tasks? a study of web navigation. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–22, 2006.

[4] A. Hassan, Y. Song, and L.-w. He. A task level metric for measuring web search satisfaction and its application on improving relevance estimation. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 125–134. ACM, 2011.

[5] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 699–708. ACM, 2008.

[6] J. Liu and C. S. Kim. Why do users perceive search tasks as difficult? exploring difficulty in different task types. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, page 5. ACM, 2013.

[7] J. Liu, C. Liu, M. Cole, N. J. Belkin, and X. Zhang. Exploring and predicting search task difficulty. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1313–1322. ACM, 2012.

[8] J. Teevan, E. Adar, R. Jones, and M. A. Potts. Information re-retrieval: repeat queries in yahoo's logs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 151–158. ACM, 2007.

[9] S. K. Tyler and J. Teevan. Large scale query log analysis of re-finding. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 191–200. ACM, 2010.

[10] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management*, 36(5):697–716, 2000.

[11] W. Webber, P. Chandar, and B. Carterette. Alternative assessor disagreement and retrieval depth. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 125–134. ACM, 2012.

[12] W. Webber, B. Toth, and M. Desamito. Effect of written instructions on assessor agreement. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1053–1054. ACM, 2012.

[13] W. V. Zhang, Y. Chen, M. Gupta, S. Sett, and T. W. Yan. Modeling click and relevance relationship for sponsored search. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 119–120. International World Wide Web Conferences Steering Committee, 2013.