A Preference Judgment Tool for Authoritative Assessment

Mahsa Seifikar
mahsa.seifikar@uwaterloo.ca
University of Waterloo, Canada

Linh Nhi Phan Minh
lnphanmi@uwaterloo.ca
University of Waterloo, Canada

Negar Arabzadeh
narabzad@uwaterloo.ca
University of Waterloo, Canada

Charles L. A. Clarke
claclark@gmail.com
University of Waterloo, Canada

Mark D. Smucker
mark.smucker@uwaterloo.ca
University of Waterloo, Canada

ABSTRACT

Preference judgments have been established as an effective method for offline evaluation of information retrieval systems with advantages to graded or binary relevance judgments. Graded judgments assign each document a pre-defined grade level, while preference judgments involve assessing a pair of items presented side by side and indicating which is better. However, leveraging preference judgments may require a more extensive number of judgments, and there are limitations in terms of evaluation measures. In this study, we present a new preference judgment tool called JUDGO, designed for expert assessors and researchers. The tool is supported by a new heap-like preference judgment algorithm that assumes transitivity and allows for ties. An earlier version of the tool was employed by NIST to determine up to the top-10 best items for each of the 38 topics for the TREC 2022 Health Misinformation track, with over 2,200 judgments collected. The current version has been applied in a separate research study to collect almost 10,000 judgments, with multiple assessors completing each topic. The code and resources are available at https://judgo-system.github.io.

CCS CONCEPTS

• **Information systems → Relevance assessment**.

KEYWORDS

offline evaluation, relevance judgment, pairwise preference

ACM Reference Format:
Mahsa Seifikar, Linh Nhi Phan Minh, Negar Arabzadeh, Charles L. A. Clarke, and Mark D. Smucker. 2023. A Preference Judgment Tool for Authoritative Assessment. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3591801

1 INTRODUCTION

Offline evaluation has been widely employed for measuring the effectiveness of Information Retrieval and Recommender systems [7, 8, 18, 34, 35]. Offline evaluation depends on gold standard labels for a set of queries, often collected through human relevance assessment [33]. To collect relevance labels, assessors are asked to assign a level of relevance to each document based on a given query or topic. Levels may be binary ("Relevant" or "Not relevant") or graded ("Perfect", "Highly relevant", etc.). For offline evaluation, relevance may depend on topical similarity and other factors such as authority, quality, and reliability [24].

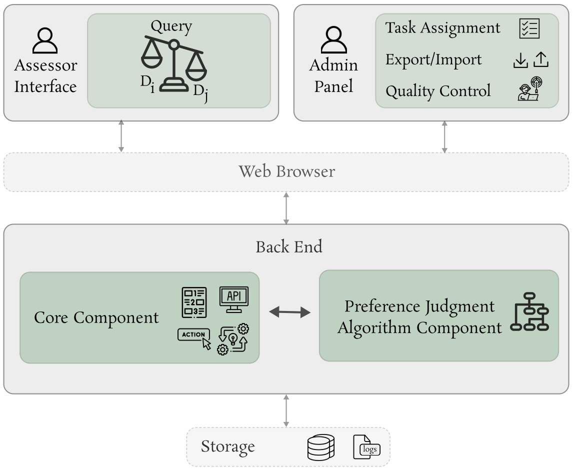


Figure 1: Overview of the preference judging tool.

Preference judgments have emerged as an alternative to traditionally graded judgment, where assessors are presented with two separate documents side-by-side and are asked to determine which one is more relevant to a given query [32]. Compared with graded judgments, preference judgments can be faster, have higher agreement among assessors, and provide better quality [9].

There are two drawbacks to preference judgments. First, that they demand more effort and are more labour-intensive, requiring around $O(NlogN)$ judgments for N documents, even under an assumption of transitivity. Second, there are no universally accepted evaluation measures for preference judgments. In recent years, research work has focused on developing evaluation metrics for preference judgments [9, 14, 33] and effective strategies to reduce the number of judgments [15, 27]. This work assumes that preference judgments will be collected on crowdsourcing websites like Amazon Mechanical Turk, where assessors are untrained and labels are noisy [2, 37].

In this paper, and its accompanying demo, we introduce JUDGO, a new preference judgment tool for authoritative assessment, which assumes that assessors are trained and motivated to make accurate assessments. Under the assumption of authoritative assessment,

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591801

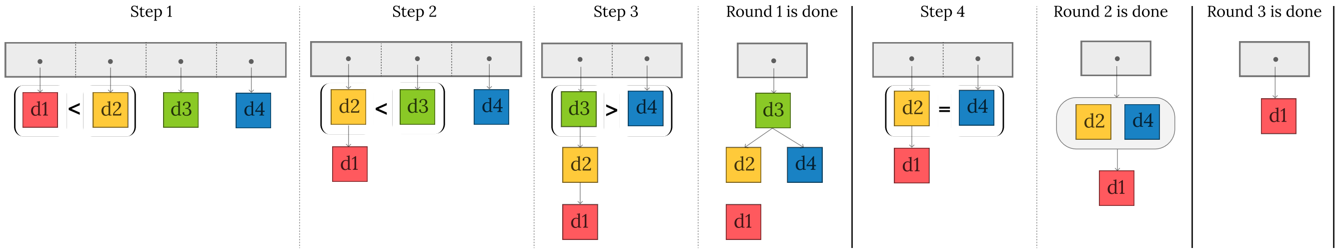


Figure 2: An example of the preference judgment algorithm with four documents. Judgment proceeds in three rounds, producing a final preference ranking of $d_1 < d_2 = d_4 < d_3$. To minimize the time spent reviewing previously seen documents, a "best known" document (or equivalence class) is maintained until a better document is identified or the round ends.

JUDGO minimizes the total number of judgments required, focusing assessment effort on identifying the top results [15]. JUDGO provides flexibility, with a separate administration interface for configuration, allowing administrators to create assessment tasks, and customize the settings and parameters to fit their requirements. In addition, JUDGO includes multiple features intended to support assessors in reading and assessing quickly and efficiently, including tags, a search box, and a progress bar.

The tool employs a tournament-like preference judgment algorithm that determines the top-k documents based on assessor preferences and transitivity assumptions. An assessor may label two documents as equally preferred, which places them in an equivalence class. To reduce the need for the assessor to re-read previously seen documents, a single "best known" document is maintained on the left, until a better document is identified, or that document, along with its equivalence class, is assessed to be the top document in the current judging round.

2 RELATED WORK

Many early IR experiments employed binary relevance judgments, where documents are labelled as either "relevant" or "not relevant" [6, 13]. Under this approach, no distinction is made between documents in the relevant set, and they are effectively considered equally relevant for measurement purposes [4]. In later experiments graded judgments became an accepted alternative, which addresses this limitation of binary judgments, with higher grades reflecting greater relevance [22, 23]. However, graded relevance judgments lack universally accepted guidelines for defining the number and interpretation of relevance grades.

Under binary or graded assessment, documents are judged one-by-one. Under preference assessment, pairs of documents are compared with one another, with the assessor choosing the most preferred or determining that they are equal. Kim et al. [24] suggest that preference judgment can incorporate a range of factors beyond topical relevance, including authority, diversity, quality, and freshness. Preference judgments could potentially be extracted from online signals, such as clicks, as well as human judgments, potentially unifying online and offline evaluation methodologies [5, 16, 19, 21, 36, 39]. For example, if a user clicks on a document, it might be assumed that the user prefers that document over all documents that are ranked higher [11, 20, 33].

Prior work often focuses on evaluation measures for preference judgments, while the preference judgment algorithm and the process itself receive relatively less attention [37]. Several strategies such as tournament-like approaches [15], sorting algorithms [9, 30, 35], active learning [31], and classifiers [17], have been proposed for extracting top items in a pool of items to judge and reducing the total number of judgments. A recent proposal by Yan et al. [37] views human preference judgments as a duelling bandit's problem. Through a series of simulations and experiments, they suggest that an improved version of one algorithm [15] was the most promising candidate for human preference judging.

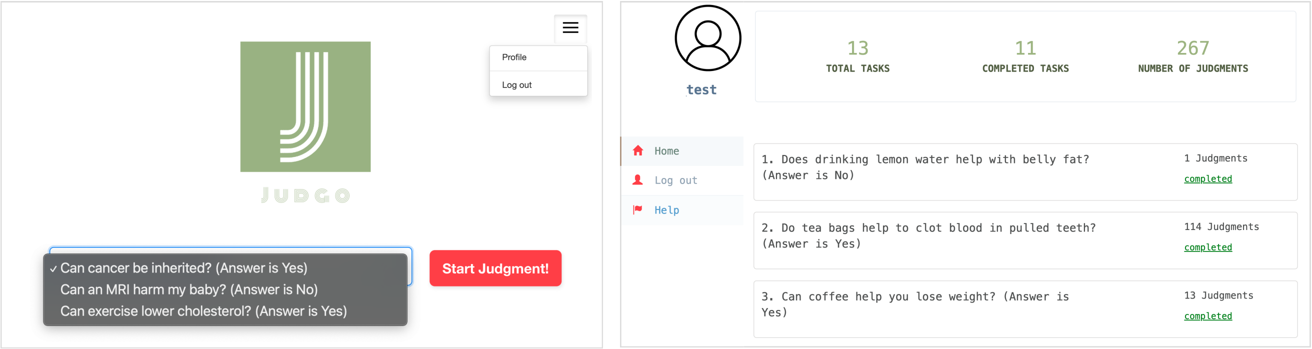
While some studies employed a small number of expert assessors, many researchers have investigated the possibility of using crowdsourcing to collect preference judgments [1, 3, 26, 28]. Yang et al. [38] reported that crowdsourced and professional assessors produced similar outcomes. They created a collection to reduce tied through the aggregation of preference, binary, and ratio assessment preferences in crowdsourced judgments.

The majority of past studies created their own private tools to meet their individual requirements [25, 38]. Carterette et al. [9] developed an interface that displays documents with their URLs and highlighted query terms to find relevant content quickly. Chandar and Carterette [10] enhanced their interface with a progress bar and topic description and concluded that assessors prefer shorter documents with fewer highlighted terms. Li et al. [27] proposed an interface with a budget monitoring panel and a sliding bar to weigh the preference between two items.

JUDGO builds on both this prior work and our own experience with preference judgment. It is an open-source tool — under ongoing development — which allows researchers to customize its settings to their specific data collection requirements. It supports a variety of features intended to facilitate assessment on varying devices — with different screen sizes — including highlighting, tags, search, font control, and sliding panels[29]. In presenting this demo, we hope to receive additional feedback on the tool, as well as providing members of the IR community with an opportunity to try preference judging for themselves on a variety of example tasks.

3 OVERVIEW OF THE INTERFACE

The JUDGO tool has three main modules, as shown in Figure 1: Backend, Assessor interface, and Admin panel. The Backend module has two components, the preference judgment and core components, which are responsible for managing the algorithm for preference judgment, interacting with the database, coordinating users' actions, and handling users and tasks. The Assessor interface has three primary pages, including the home page, profile page, and judgment page. The admin panel has three key components: task assignment, quality control module, and import/export of information.



✓ Can cancer be inherited? (Answer is Yes)
Can an MRI harm my baby? (Answer is No)
Can exercise lower cholesterol? (Answer is Yes)

test

13 TOTAL TASKS

11 COMPLETED TASKS

267 NUMBER OF JUDGMENTS

1. Does drinking lemon water help with belly fat? (Answer is No) — 1 Judgments — completed

2. Do tea bags help to clot blood in pulled teeth? (Answer is Yes) — 114 Judgments — completed

3. Can coffee help you lose weight? (Answer is Yes) — 13 Judgments — completed

A) Home Page

B) Profile Page

Figure 3: The assessor interface in the JUDGO tool; A) Home Page presents users with assigned topics. B) Profile Page displays the assessor's progress and past activities.

We have designed a database with five tables for storing information on assessors, topics, documents, tasks, and judgments. The backend directly interacts with the database and passes data through web browsers to the assessor interface and admin panel. The tool is intended for two types of users: Admin researchers who initiate the judgments and assessors who conduct the judgments. The former interacts with the admin panel and the latter works with the assessor interface.

3.1 The Preference Judgment Algorithm

The preference judgment algorithm, running in the backend of the JUDGO tool is responsible for ranking documents based on user actions, utilizing an algorithm that maintains judging state with a heap-like data structure. This algorithm assumes transitivity and follows a tournament-style approach with multiple rounds. During each round, one or more documents are selected as winners according to the assessor's preferences, and placed in a separate relevance level. Additionally, the algorithm transforms equal items into equivalence classes. The heap-like data structure maintains the priority of items, where the data structure recursively consists of a top item (or equivalence class) and a list of children that are less preferable than the top item. The algorithm requires two inputs; a pool of documents and a threshold that determines the minimum number of top documents to be retrieved before stopping.

Figure 2 shows a running example of preference judgment algorithm with four documents $d_1 - d_4$. In the initial step (Step 1), a list with four heap-like data structures, each containing a single document is created from a set of four documents selected for a given topic (query). The algorithm started by removing the first two elements from the list, extracting their top item and presenting them side-by-side to the assessor for assessment.

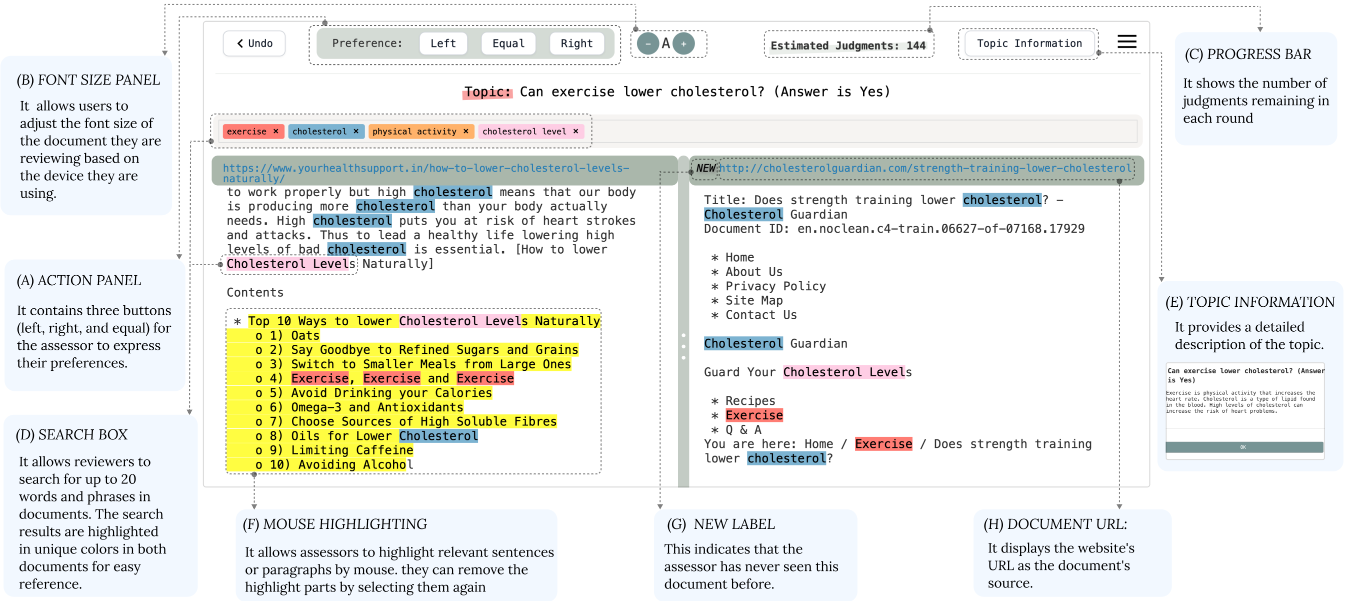
In step one of Figure 2, the assessor is presented with d_1 and d_2 as the left and right documents in the JUDGO assessor interface. Assuming the assessor prefers document d_2 over d_1, a new heap-like data structure is created with d_2 on the top and d_1 as its child. This new structure is added to the beginning of the list as shown in step 2 of Figure 2. As such, the assessor is exposed to only one new document (d_3) in step 2, as the algorithm takes advantage of the assessor's previous decision and shows them the preferred document (d_2) from the previous step along with the new document (d_3), avoiding the need to re-read the preferred document.

The assessor prefers d_3 over d_2 in step 2 and subsequently selects d_3 over d_4 in step 3. The first round of the algorithm is over. At this point, there is only one heap-like data structure item left on the list, and its top item (d_3) is considered as the winner of the round. Thus, in the first round, d_3 is determined to be in the top rank. When one round of the algorithm is completed, the single item in the list is popped and its children are added to the list as shown in step 4 of Figure 2. In step 4 of Figure 2, the assessor determines that documents d_2 and d_4 are equally preferred. A new heap-like data structure is created with both documents placed in an equivalence node shown as an oval shape in Figure 2. This equivalence node will be considered as the top item of this new heap-like data structure. As a result, both documents are regarded as winners of the second round and form the second-best relevance level. Since only document d_1 remains on the list, it forms a third relevance level. The algorithm then terminates since there are no additional items in the list to assess.

3.2 The Assessor Interface

When assessors log in to the JUDGO tool, they are directed to the homepage, as shown in Figure 3 (A), where they are presented with a list of topics. The assessors choose from the list, allowing them to decide what task to work on next, if several are available. In the top right corner, there is a menu that includes the user's profile and logout. The profile page provides an overview of the user's activities, including the total number of tasks, completed tasks, and total judgments performed, as well as a list of assigned tasks with their current state and the total number of required judgments.

Once a user clicks on "Start Judgment!" button, they are directed to the judgment page. As shown in Figure 4, the judgment page includes a toolbar and the two documents sections which are separated by a drag bar. Each document has an associated URL (H), title, and content. Assessors are provided with a unique pair of documents to assess at each step of the review process, and the *"NEW"* label (G) is displayed on the top of any document which has not been seen before. Assessors are required to read both documents, make a side-by-side comparison, and use the action panel (A) to determine which document, either the one on the left or right is more relevant to the given topic. The "equal" button could be used in case the contents of the documents are identical or they have the same level of preference.



Topic: Can exercise lower cholesterol? (Answer is Yes)

https://www.yourhealthsupport.in/how-to-lower-cholesterol-levels-naturally/

to work properly but high cholesterol means that our body is producing more cholesterol than your body actually needs. High cholesterol puts you at risk of heart strokes and attacks. Thus to lead a healthy life lowering high levels of bad cholesterol is essential. [How to lower Cholesterol Levels Naturally]

Contents

```
* Top 10 Ways to lower Cholesterol Levels Naturally
    o 1) Oats
    o 2) Say Goodbye to Refined Sugars and Grains
    o 3) Switch to Smaller Meals from Large Ones
    o 4) Exercise, Exercise and Exercise
    o 5) Avoid Drinking your Calories
    o 6) Omega-3 and Antioxidants
    o 7) Choose Sources of High Soluble Fibres
    o 8) Oils for Lower Cholesterol
    o 9) Limiting Caffeine
    o 10) Avoiding Alcohol
```

http://cholesterolguardian.com/strength-training-lower-cholesterol

Title: Does strength training lower cholesterol? – Cholesterol Guardian
Document ID: en.noclean.c4-train.06627-of-07168.17929

```
 * Home
 * About Us
 * Privacy Policy
 * Site Map
 * Contact Us
```

Cholesterol Guardian

Guard Your Cholesterol Levels

```
 * Recipes
 * Exercise
 * Q & A
```

You are here: Home / Exercise / Does strength training lower cholesterol?

(B) FONT SIZE PANEL

It allows users to adjust the font size of the document they are reviewing based on the device they are using.

(A) ACTION PANEL

It contains three buttons (left, right, and equal) for the assessor to express their preferences.

(D) SEARCH BOX

It allows reviewers to search for up to 20 words and phrases in documents. The search results are highlighted in unique colors in both documents for easy reference.

(F) MOUSE HIGHLIGHTING

It allows assessors to highlight relevant sentences or paragraphs by mouse. they can remove the highlight parts by selecting them again

(G) NEW LABEL

This indicates that the assessor has never seen this document before.

(H) DOCUMENT URL:

It displays the website's URL as the document's source.

(C) PROGRESS BAR

It shows the number of judgments remaining in each round

(E) TOPIC INFORMATION

It provides a detailed description of the topic.

Can exercise lower cholesterol? (Answer is Yes)

Exercise is physical activity that increases the heart rate. Cholesterol is a type of lipid found in the blood. High levels of cholesterol can increase the risk of heart problems.

Figure 4: The assessor interface in the JUDGO tool; Judgment page presents a user with two documents side-by-side and provides them with various features for facilitating the decision-making process.

The tool also includes several features that assist assessors in evaluating pairs of documents and selecting the preferred one, such as the topic information button (E) which helps to understand the information need behind the query, especially when the assessor suffers from a lack of expertise on the topic. A font change panel (B) and dragbar accommodate different screen sizes; an undo button allows assessors to go back and revise their previous decisions. In addition, there is a progress bar (C) which indicates the number of judgments left to complete until the next round of the algorithm. The assessors can log out or return to the homepage to work on other tasks at any point during the judging process.

In order to facilitate assessment, a search box (D) and highlighting option (F) are provided for assessors potentially speeding up the process of reading and decision-making. Specified search terms are highlighted in different colours on the document to attract the attention of the assessors as shown in Figure 4. Moreover, the highlighted parts serve as a quick reference for the assessors, allowing them to focus on the critical parts of the document and make a judgment more accurately and effectively.

3.3 The Admin Panel

The first component of the admin panel – export/import — uploads lists of assessors and tasks, and downloads the final assessment results to a CSV file. The second component — task assignment — enables the administrator to assign topics to assessors. In addition to manual assignments, they can also upload a list of tasks using the importing tool. The quality control component randomly selects a pair of previously judged documents to measure consistency. If the ratio of consistent tests to total tests falls below a predefined threshold, an informational message can be sent.

4 EXPERIENCE

An earlier version of the tool was used by NIST to determine up to the top-10 best items for each of 38 topics for the TREC 2022 Health Misinformation Track, with 2,200 judgments collected [12]. The current version of the tool reflects feedback received from NIST, including bug reports. Based on this feedback, we made several changes reflected in the version described in this paper. In particular, earlier versions of the tool did not maintain the current "best document", which increased assessor effort by requiring them to re-review a greater number of previously seen documents. Other changes included highlighting features and more robust logging.

More recently, we have conducted additional experiments to re-judge topics from the TREC 2021 Health Misinformation Track to explore assessor consistency. In this second experiment, we employed 40 assessors to evaluate 30 topics, with each topic being evaluated by three assessors. During this study, we collected almost 10,000 additional judgments, with more than 300 words or phrases entered in the search box and roughly 20,460 parts of 1,623 documents highlighted by the mouse. We are continuing to analyze this data, which will be reported in a forthcoming publication.

5 CONCLUSION

Many SIGIR 2023 attendees may not have had personal experience with preference judgments, and this demo provides an opportunity for attendees to explore them and discuss their potential. The demo will include a variety of examples taken from the TREC Health Misinformation Track, the TREC Deep Learning Track, and the MS Marco collection. In conducting the demo, we hope to receive additional feedback, allowing us to continue to improve the tool.

ACKNOWLEDGMENTS

We thank Ehsan Jahangirzadeh Soure for his feedback. This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada (RGPIN-2020-04665, RGPAS2020-00080) and in part by the Digital Research Alliance of Canada.

REFERENCES

[1] Omar Alonso, Daniel E Rose, and Benjamin Stewart. 2008. Crowdsourcing for relevance evaluation. In *ACM SigIR forum*, Vol. 42. ACM New York, NY, USA, 9–15.

[2] Negar Arabzadeh, Alexandra Vtyurina, Xinyi Yan, and Charles LA Clarke. 2022. Shallow pooling for sparse labels. *Information Retrieval Journal* 25, 4 (2022), 365–385.

[3] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P de Vries, and Emine Yilmaz. 2008. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.* 667–674.

[4] Maryam Bashir, Jesse Anderton, Jie Wu, Peter B Golbus, Virgil Pavlu, and Javed A Aslam. 2013. A document rating system for preference judgements. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval.* 909–912.

[5] Mike Bendersky, Xuanhui Wang, Marc Najork, and Don Metzler. 2018. Learning with Sparse and Biased Feedback for Personal Search. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI).* 5219–5223.

[6] Pia Borlund. 2003. The concept of relevance in IR. *Journal of the American Society for information Science and Technology* 54, 10 (2003), 913–925.

[7] Rocío Cañamares, Pablo Castells, and Alistair Moffat. 2020. Offline evaluation options for recommender systems. *Information Retrieval Journal* 23, 4 (2020), 387–410.

[8] Ben Carterette. 2011. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in information retrieval.* 903–912.

[9] Ben Carterette, Paul N Bennett, David Maxwell Chickering, and Susan T Dumais. 2008. Here or there. In *European Conference on Information Retrieval.* Springer, 16–27.

[10] Praveen Chandar and Ben Carterette. 2012. Using preference judgments for novel document retrieval. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval.* 861–870.

[11] Charles LA Clarke, Chengxi Luo, and Mark D Smucker. 2021. Evaluation measures based on preference graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1534–1543.

[12] Charles LA Clarke, Maria Maistro, Mahsa Seifikar, and Mark D Smucker. 2022. Overview of the TREC 2022 health misinformation track. In *TREC*.

[13] Charles LA Clarke, Mark D Smucker, and Alexandra Vtyurina. 2020. Offline evaluation by maximum similarity to an ideal ranking. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management.* 225–234.

[14] Charles LA Clarke, Alexandra Vtyurina, and Mark D Smucker. 2020. Offline evaluation without gain. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval.* 185–192.

[15] Charles LA Clarke, Alexandra Vtyurina, and Mark D Smucker. 2021. Assessing Top-k Preferences. *ACM Transactions on Information Systems (TOIS)* 39, 3 (2021), 1–21.

[16] Zhicheng Dou, Ruihua Song, Xiaojie Yuan, and Ji-Rong Wen. 2008. Are click-through data adequate for learning web search rankings?. In *Proceedings of the 17th ACM conference on Information and knowledge management.* 73–82.

[17] Ahmed Hassan Awadallah and Imed Zitouni. 2014. Machine-assisted search preference evaluation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management.* 51–60.

[18] Kai Hui and Klaus Berberich. 2017. Transitivity, time consumption, and quality of preference judgments in crowdsourcing. In *European Conference on Information Retrieval.* Springer, 239–251.

[19] Thorsten Joachims. 2003. Evaluating Retrieval Performance Using Clickthrough Data. In *Text Mining*. Physica/Springer Verlag, 79–96.

[20] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)* 25, 2 (2007), 7–es.

[21] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the tenth ACM international conference on web search and data mining.* 781–789.

[22] Jaana Kekäläinen. 2005. Binary and graded relevance in IR evaluations—comparison of the effects on ranking of IR systems. *Information processing & management* 41, 5 (2005), 1019–1033.

[23] Jaana Kekäläinen and Kalervo Järvelin. 2002. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology* 53, 13 (2002), 1120–1129.

[24] Jinyoung Kim, Gabriella Kazai, and Imed Zitouni. 2013. Relevance dimensions in preference-based IR evaluation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval.* 913–916.

[25] Caitlin Kuhlman, Diana Doherty, Malika Nurbekova, Goutham Deva, Zarni Phyo, Paul-Henry Schoenhagen, MaryAnn VanValkenburg, Elke Rundensteiner, and Lane Harrison. 2019. Evaluating preference collection methods for interactive ranking analytics. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–11.

[26] Matthew Lease and Emine Yilmaz. 2012. Crowdsourcing for information retrieval. In *ACM SIGIR Forum*, Vol. 45. ACM New York, NY, USA, 66–75.

[27] Yan Li, Hao Wang, Ngai Meng Kou, Zhiguo Gong, et al. 2021. Crowdsourced top-k queries by pairwise preference judgments with confidence and budget control. *The VLDB Journal* 30, 2 (2021), 189–213.

[28] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. 2017. On crowdsourcing relevance magnitudes for information retrieval evaluation. *ACM Transactions on Information Systems (TOIS)* 35, 3 (2017), 1–32.

[29] Mariana Neves and Jurica Ševa. 2021. An extensive review of tools for manual annotation of documents. *Briefings in bioinformatics* 22, 1 (2021), 146–163.

[30] Shuzi Niu, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2012. Top-k learning to rank: labeling, ranking and evaluation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval.* 751–760.

[31] Kira Radinsky and Nir Ailon. 2011. Ranking from pairs and triplets: Information quality, evaluation methods and query complexity. In *Proceedings of the fourth ACM international conference on Web search and data mining.* 105–114.

[32] Kevin Roitero, Alessandro Checco, Stefano Mizzaro, and Gianluca Demartini. 2022. Preferences on a Budget: Prioritizing Document Pairs when Crowdsourcing Relevance Judgments. In *Proceedings of the ACM Web Conference 2022.* 319–327.

[33] Tetsuya Sakai and Zhaohao Zeng. 2020. Good evaluation measures based on document preferences. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 359–368.

[34] Mark Sanderson and Justin Zobel. 2005. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval.* 162–169.

[35] Ruihua Song, Qingwei Guo, Ruochi Zhang, Guomao Xin, Ji-Rong Wen, Yong Yu, and Hsiao-Wuen Hon. 2011. Select-the-Best-Ones: A new way to judge relative relevance. *Information processing & management* 47, 1 (2011), 37–52.

[36] Paul Thomas and David Hawking. 2006. Evaluation by comparing result sets in context. In *Proceedings of the 15th ACM international conference on Information and knowledge management.* 94–101.

[37] Xinyi Yan, Chengxi Luo, Charles LA Clarke, Nick Craswell, Ellen M Voorhees, and Pablo Castells. 2022. Human preferences as dueling bandits. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 567–577.

[38] Ziying Yang, Alistair Moffat, and Andrew Turpin. 2018. Pairwise crowd judgments: Preference, absolute, and ratio. In *Proceedings of the 23rd Australasian Document Computing Symposium.* 1–8.

[39] Zhaohui Zheng, Keke Chen, Gordon Sun, and Hongyuan Zha. 2007. A regression framework for learning ranking functions using relative relevance judgments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.* 287–294.