Understanding and Predicting Usefulness Judgment in Web Search

Jiaxin Mao[†], Yiqun Liu[†*], Huanbo Luan[†], Min Zhang[†],
Shaoping Ma[†], Hengliang Luo[‡], Yuntao Zhang[‡]
[†]Department of Computer Science & Technology, Tsinghua University, Beijing, China
[‡]Samsung Beijing R&D Center, Beijing, China
yiqunliu@tsinghua.edu.cn

ABSTRACT

Usefulness judgment measures the user-perceived amount of useful information for the search task in the current search context. Understanding and predicting usefulness judgment are crucial for developing user-centric evaluation methods and providing contextualize results according to the search context. With a dataset collected in a laboratory user study, we systematically investigate the effects of a variety of content, context, and behavior factors on usefulness judgments and find that while user behavior factors are most important in determining usefulness judgments, content and context factors also have significant effects on it. We further adopt these factors as features to build prediction models for usefulness judgments. An AUC score of 0.909 in binary usefulness classification and a Pearson's correlation coefficient of 0.694 in usefulness regression demonstrate the effectiveness of our models. Our study sheds light on the understanding of the dynamics of the user-perceived usefulness of documents in a search session and provides implications for the evaluation and design of Web search engines.

CCS CONCEPTS

•**Information systems → Users and interactive retrieval;** *Retrieval effectiveness;*

KEYWORDS

Usefulness, User Behavior Analysis, Evaluation

1 INTRODUCTION

Web search engines help people effectively deal with the information overload by retrieving a small number of highly relevant documents to the user within a second. However, high relevance (especially topical relevance) between the document and query may not necessarily mean the document is useful for the user [10]. It is still very common for the user to encounter documents with low usefulness and feel frustrated in the search. This gap motivates us to study the dynamic, situational, and subjective document-level *usefulness judgment*, which is defined as the user-perceived amount of useful information in the document, for the search task at hand, in the current search context. From the perspective of search engine evaluation, investigating usefulness judgment can help us better understand user's search process and develop user-centric evaluation methods. From the perspective of search engine design, usefulness judgments from real users can be used as signals of high-quality results and the prediction of usefulness judgment may guide the system to provide personalized and contextualized results for the user.

In order to study user's usefulness judgment, we conducted a user study in laboratory settings to collect a dataset that contains fine-grain search logs for 166 sessions and the corresponding usefulness judgments from users on 1,383 visited documents. Using this dataset, we try to answer the following research questions:
RQ1: What factors may affect users' usefulness judgments?
RQ2: Can we estimate or predict usefulness judgments using such factors as features?

To address **RQ1**, we examine the effect of *content*, *context*, and *behavior* factors on usefulness judgments. To address **RQ2**, we use these factors as features to build regression and (binary) classification models to estimate and predict usefulness judgments. The prediction performance is promising in that the classification model achieves an AUC score of 0.909 and the Pearson's r between actual usefulness judgments and the prediction of the regression model reaches 0.694.

Related Work

Some recent studies have addressed concept and measurement of usefulness in information retrieval.

Belkin et al. [1, 2] proposed to adopt usefulness in evaluating interactive information retrieval systems. Mao et al. [10] found that topical relevance is necessary but not sufficient for usefulness and the usefulness judgment correlates better with user satisfaction than relevance judgment. Kim et al. [7] used an online experiment to collect users' in situ feedback of whether a search result is helpful or not. They found that the in situ usefulness feedback is different from assessors' relevance judgment because the user may have idiosyncratic search intents and the ideal threshold of dwell time for predicting positive in situ feedbacks is much longer than the ideal threshold for predicting positive relevance judgments (87 s vs. 38 s). Jiang et al. [6] analyzed the relationship between contextual usefulness feedback(called *ephemeral state of relevance (ESR)* in

*Corresponding author

SIGIR '17, Shinjuku, Tokyo, Japan

DOI: http://dx.doi.org/10.1145/3077136.3080750

their paper) and other explicit feedbacks such as topical relevance, novelty, effort, understandability, and reliability.

Compared to these studies, our work focuses on inspecting a variety *implicit* factors that may affect usefulness judgment as well as building a prediction model that utilizes these implicit factors to effectively estimate user's usefulness judgments.

2 DATA COLLECTION

In this section, we describe the settings of the user study and the dataset we collected.

The procedure of the user study is shown in Figure 1. In the user study, we required each participant complete 6 search tasks[1] using an experiment search engine. A total of 30 participants were recruited via email or online social networks. All of them are college students aged from 19 to 22. 22 participants are female, and other 8 participants are male. Each search task is a non-trivial question that can be answered in 100 words such as: *What are the most commonly-used treatments for cancer in clinical?*. Before carrying out the first search task, each participant would go through a pre-experiment training stage. Because we also recorded participants' eye fixation movements using a Tobii X2-30 eye-tracker, in stage II, we calibrated the eye-tracker for each participant.

For each of the six search tasks, the participant was required to go through stage III to stage VII. First, in stage III, the participant would read and remember the task description (i.e. a question) on a web page. We further required the participant re-input the task description on the next web page to make sure he or she actually remember the search task. After that, in stage IV. pre-task questionnaire we collected participants' expected difficulty, interest, and prior domain knowledge level about the search task in 5-point Likert scales. Then in stage V, the participant would use an experiment search engine to gather information to answer the question in the task description. The experiment search engine has a similar interface as a common commercial search engine, and the search results are crawled from Bing in real time when receive queries from the participants. We used a Chrome extension to log participants' querying, clicking, tab-switch, scrolling, and mouse movement actions during the search. To collect usefulness judgment, we instructed the participant to use a group of radio buttons in the right-click menu, which is injected by the Chrome extension, to annotate useful documents in the search trail. The instruction and scale for the usefulness judgment are the following:

Is the document useful for the completion of your search tasks? If it is, please use the right-click menu to annotate the usefulness on the web page in the following scale.

1: not useful at all; 2: somewhat useful
3: fairly useful; 4: very useful

The default option is "1: not useful at all", so the participant only needed to annotate the documents that are at least "2: somewhat useful". After completing the search task, in stage VI, the participant would answer the question in the task description and the answer would be logged by the experiment system. Finally, in stage VII, the participant would give feedbacks about the perceived difficulty, interest, knowledge gain, and satisfaction level in a post-task questionnaire.

[1]The search tasks were selected from our previous work[9].

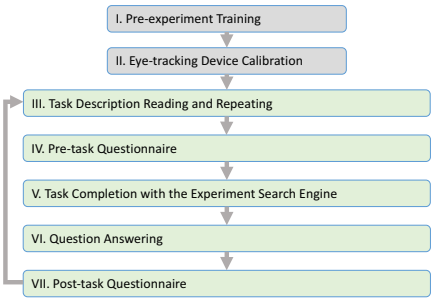


Figure 1: The procedure of the user study.

After filtering the search sessions in which the eye-tracker malfunctioned, we collected dataset that contains 166 valid search sessions from 28 unique participants. A total of 1,383 documents were visited. Among these documents, 897 (64.86%) were not annotated or judged as "1: not useful at all", 184 (13.30%) as "2: somewhat useful", 172 (12.44%) as "3: fairly useful", and 130 (9.40%) as "4: very useful". The participant visited 2.93 useful documents (at least "2: somewhat useful") per session and the average usefulness judgment is 1.66.

3 FACTOR ANALYSIS ON USEFULNESS JUDGMENTS

In this section, we examine three groups of factors: *content* factors, *context* factors, and *behavior* factors and investigate their effects on users' usefulness judgments to answer **RQ1**.

3.1 Content Factors

The content factors include the page contents' cosine similarities with the corresponding query (content_cossim_with_query), task description (content_cossim_with_task_description), and answer (content_cossim_with_answer) as well as the Okapi BM25 score of the query-document pair (content_bm25_with_query). We also use the eye-tracking data to infer which term on the page is actually fixated by the participant. Therefore, we further compute the cosine similarities and Okapi BM25 scores based on the fixated page contents (denoted as fix_content).

We measure the effect of each content factor on usefulness judgment by the Pearson's correlation coefficient r between the factor and usefulness judgment. From the results shown in Table 1 we can see that: 1) All the content factors have significantly positive effects on usefulness judgments. 2) Among all the content factors, the cosine similarities with the answer submitted by the participant have the strongest correlations with usefulness judgments, suggesting that usefulness judgments are largely determined by whether the information on the page is useful for the completion of the search task, which happens to be answering an question in our study. 3) Compare with page contents, the fixated contents have stronger correlations with usefulness judgments.

3.2 Context Factors

The search context is determined by the search task and previous user interactions in the search session. Previous studies show that

Table 1: Effects of content factors on usefulness judgment (*//*** indicates the correlation is statistically significant at $p < 0.05/0.01/0.001$)**

Factors	Pearson's r	p
content_bm25_with_query	0.180	***
content_cossim_with_answer	0.335	***
content_cossim_with_query	0.056	0.038*
content_cossim_with_task_description	0.169	***
fix_content_bm25_with_query	0.238	***
fix_content_cossim_with_answer	0.406	***
fix_content_cossim_with_query	0.181	***
fix_content_cossim_with_task_description	0.238	***

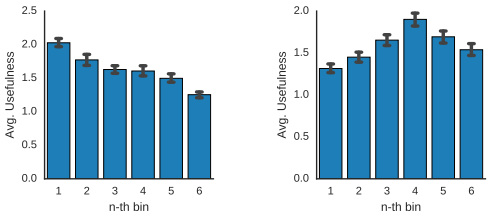
Table 2: Effects of context factors on usefulness judgment (*//*** indicates the correlation is statistically significant at $p < 0.05/0.01/0.001$)**

Factors	Pearson's r	p
avg_cossim_with_previous_page_content	0.020	0.467
max_cossim_with_previous_page_content	-0.052	0.052
avg_usefulness_of_previous_page	-0.024	0.378
total_usefulness_of_previous_page	-0.206	***
max_usefulness_of_previous_page	-0.161	***
num_previous_doc	-0.214	***
num_previous_doc_in_query	-0.163	***
num_previous_query	-0.100	***
progress_time_in_session	-0.159	***
task_difficulty	-0.079	***
task_domain_knowledge	-0.025	0.348
task_interest	0.082	0.002**

the search context factors can explain why the usefulness judgment of a document is different from the relevance of it [3]. We extract context factors based on feedbacks in pre-task questionnaires and user's previous behavior in session and analyze their influence on usefulness judgments.

We show the effects of context factors in Table 2. From the results, we find that the usefulness judgment tends to decrease as the search proceed. A number of context factors such as the number of previously visited documents (num_previous_doc), the sum of the usefulness judgments of previous documents (total_usefulness_of_previous_page), and the time spent in session (progress_time_in_session) have negative correlations with usefulness. We further show the interaction between num_previous_doc and usefulness in Figure 2a. These findings suggest that the usefulness judgment measures the increment of useful information when visiting a document, which is likely to diminish as the progress of searching.

This diminishing return may be caused by the redundancy with previous documents. We use the cosine similarity with previously visited documents to capture the redundancy factor and investigate its relationship with usefulness judgments. A non-monotonous relationship between max_cossim_with_previous_page_content and usefulness is spotted. As shown in Figure 2b, the usefulness first increase with the max cosine similarity and then decrease with it. When a document is not similar to any of previously visited documents, it is likely to be irrelevant to user's current information need. But if it is very similar to one of the visited documents, it will be redundant to the visited one thus not useful for the user.



(a) (b)

Figure 2: The effect of two context factors: (a) the number of visited documents (num_previous_doc); (b) the max cosine similarity between current document and previous visited documents (max_cossim_with_previous_page_content). To show these effects, we put the documents into 6 bins with equal size according to the factor and compute average usefulness for each bin.

Table 3: Effects of user behavior factors on usefulness judgment (*//*** indicates the correlation is statistically significant at $p < 0.05/0.01/0.001$)**

Factors	Pearson's r	p
avg_eye_speed	-0.109	***
avg_mouse_movement_speed	-0.098	***
fixation_num	0.582	***
mouse_movement_num	0.519	***
page_dwell_time	0.568	***
query_length	0.042	0.116
query_time	0.008	0.756
scroll_num	0.430	***
session_time	-0.095	***
viewport_coverage	0.144	***

3.3 Behavior Factors

We also inspect a variety of behavior factors that characterize user's actions on the document.

From the results in Table 3, we first find that, similar to the results reported in existing work [7, 8], the dwell time on page (page_dwell_time) is strongly correlated with usefulness. Because the user tends to stay longer on useful documents, other measures of user engagement such as the number of eye fixations (fixation_num), the number of mouse movements (mouse_movement_num), and the number of scrolling actions (scroll_num) also show positive correlations with usefulness judgments. We also find that the average moving speed of the eye fixation point and mouse cursor is negatively correlated with usefulness judgment. This confirms with Buscher et al. [4]'s finding that users prefer skimming to reading when accessing irrelevant documents.

4 USEFULNESS PREDICTION

After inspecting the effects of content, context, and behavior factors on usefulness judgments, we try to utilize them as features to build prediction models for the user-perceived usefulness.

We adopt two experiment settings for usefulness prediction: *binary usefulness classification* and *usefulness regression*. For *binary usefulness classification*, we build a classification model to predict a

Table 4: Results of usefulness prediction. */ indicates the difference with the prediction result use all the features (All) is significant at $p < 0.05/0.01$.**

	Binary Usefulness Classification			Usefulness Regression		
	AUC	Accuracy	F1	Pearson's r	MSE	MAE
content	0.753±0.012**	0.698±0.006**	0.545±0.019**	0.431±0.024**	0.845±0.031**	0.703±0.015**
context	0.718±0.014**	0.682±0.006**	0.494±0.022**	0.299±0.026**	0.947±0.038**	0.779±0.017**
behavior	0.905±0.008	0.833±0.004*	0.764±0.014*	0.679±0.020*	0.559±0.026*	0.493±0.015
content+context	0.782±0.012**	0.722±0.006**	0.574±0.019**	0.451±0.024**	0.827±0.031**	0.679±0.016**
content+behavior	0.904±0.008*	0.836±0.004*	0.770±0.014**	0.686±0.020*	0.549±0.025*	0.489±0.015
context+behavior	0.905±0.008	0.834±0.004	0.763±0.015*	0.692±0.019	0.541±0.024	0.498±0.014
All	0.909±0.008	0.848±0.003	0.785±0.015	0.694±0.019	0.537±0.024	0.494±0.014

binary variable indicating whether a document is useful or not for the user and use Area-Under-Curve of ROC (AUC), accuracy, and the F1 score for the useful documents to evaluate its performance. Because 64.86% documents are "1: not useful at all", we use them as negative samples (not-useful documents) and other documents that are at least "2: somewhat useful" as positive samples (useful document). For usefulness regression, we build a regression model to predict the actual usefulness judgment scale (a real number ranging from 1 to 4). The evaluation metrics adopted are Pearson's r, Mean Squared Error (MSE), and Mean Absolute Error (MAE).

We adopt Gradient Boosting Tree [5] model in both binary usefulness classification and usefulness regression settings because it can handle heterogeneous features and has a good prediction power.

The results of prediction performance with different feature combinations are shown in Table 4. All the evaluation metrics were computed using a 10-fold cross-validation over sessions. From the results we can see that: 1) Using all the features, the Gradient Boosting Tree can effectively estimate usefulness judgment. The AUC in binary usefulness classification reaches 0.909 and the Pearson's r in usefulness regression is 0.694. 2) The behavior features are the most informative features in usefulness prediction. Using only the behavior features, the model can achieve an AUC of 0.905 and a Pearson's r of 0.679. Adding content and context features only slightly improves the prediction performance. 3) Given the content and context features, we can predict usefulness to a moderate extent with an AUC of 0.782 and a Pearson's r of 0.451. Because most of the content and context features can be computed *before* the user visits the page, this usefulness prediction can be used to identify the documents that are likely to be useful in the current search context.

5 CONCLUSIONS AND DISCUSSIONS

With the dataset collected in a user study, we investigate the effects of various content, context, and behavior factors on users' usefulness judgments during the search (**RQ1**), and further use these factors as features to build effective prediction models for usefulness judgments (**RQ2**). Our findings not only enriches the understanding on the dynamic, situational usefulness judgments but also provide some implications for the evaluation and design of Web search engines. For example, the findings in section 3.1 suggest that, in a user study, we can use the similarity between page contents and submitted answers to infer usefulness judgments and use them to evaluate the search performance without explicit feedbacks from users. The findings in section 3.2 suggest that, in order to avoid returning irrelevant or redundant results to users, the search system should return results that are moderately similar to previously visited documents. And the results of usefulness prediction suggest that: 1) With the behavior features in search logs, we can accurately infer the usefulness judgments. These judgments can be used as offline ranking signals and evaluation measures for the system. 2) With the content and context features that can be obtained before the user actually clicks the document, we can predict the usefulness judgment to provide contextualize result ranking for the user.

ACKNOWLEDGMENTS

This work was supported by Natural Science Foundation of China (Grant No. 61622208, 61532011, 61672311), Tsinghua University Initiative Scientific Research Program(2014Z21032), National Key Basic Research Program (2015CB358700).

REFERENCES

[1] Nicholas J Belkin. 2015. Salton award lecture: people, interacting with information. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1–2.

[2] Nicholas J Belkin, Michael Cole, and Jingjing Liu. 2009. A model for evaluation of interactive information retrieval. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*. 7–8.

[3] Alexey Borisov, Ilya Markov, Maarten de Rijke, and Pavel Serdyukov. 2016. A context-aware time model for web search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 205–214.

[4] Georg Buscher, Andreas Dengel, and Ludger van Elst. 2008. Eye movements as implicit relevance feedback. In *CHI'08 extended abstracts on Human factors in computing systems*. ACM, 2991–2996.

[5] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[6] Jiepu Jiang, Daqing He, Diane Kelly, and James Allan. 2017. Understanding Ephemeral State of Relevance. In *CHIIR'17*.

[7] Jin Young Kim, Jaime Teevan, and Nick Craswell. 2016. Explicit In Situ User Feedback for Web Search Results. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. ACM, New York, NY, USA, 829–832. DOI:http://dx.doi.org/10.1145/2911451.2914754

[8] Youngho Kim, Ahmed Hassan, Ryen W White, and Imed Zitouni. 2014. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 193–202.

[9] Xin Li, Yiqun Liu, Rongjie Cai, and Shaoping Ma. 2017. Investigation of User Search Behavior While Facing Heterogeneous Search Services. In *WSDM'17*. ACM, 161–170.

[10] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. 2016. When Does Relevance Mean Usefulness and User Satisfaction in Web Search?. In *SIGIR'16*. ACM, 463–472.