How do Users Perceive Information: Analyzing user feedback while annotating textual units

Piyush Arora
ADAPT Centre, School of Computing
Dublin City University
Ireland
parora@computing.dcu.ie

Gareth J. F. Jones
ADAPT Centre, School of Computing
Dublin City University
Ireland
gfjones@computing.dcu.ie

ABSTRACT

We describe an initial study of how participants perceive information when they categorize highlighted textual units within a document marked for a given information need. Our investigation explores how users look at different parts of the document and classify textual units within retrieved documents on 4-levels of relevance and importance. We compare how users classify different textual units within a document, and report mean and variance for different users across different topics. Further, we analyze and categorise the reasons provided by users while rating textual units within retrieved documents. This research shows some interesting observations regarding *why* some parts of the document are regarded as more relevant than others (e.g. it provides contextual information, contains background information) and *which* kind of information seems to be effective for satisfying the end users (e.g showing examples, providing facts) in a search task. This work is a part of our ongoing investigation into generation of effective surrogates and document summaries based on search topics and user interactions with information.

1 INTRODUCTION

Information retrieval (IR) focuses on optimizing topical relevance by retrieving documents that are relevant to the user's information need [7]. Initial work in IR focused on assessing documents relevance at binary levels (relevant and non-relevant items), but subsequently shifted towards more graded relevance levels (highly relevant, partially relevant, non relevant items). Work by Spink et al. [15] categorizes the prior work on user oriented relevance based on two mains aspects: 1) levels of relevance, and 2) regions of relevance. In their work the authors studied different regions of relevance and their relation to changes in the user's information problem definition, the user's personal knowledge, the searcher's intermediaries perception and the user's criteria for marking relevance judgements. They examined the criteria for marking retrieved items as relevant, partially relevant and non-relevant. Further, they proposed methods by which future search systems should support highly relevant and partially relevant items depending on the user's knowledge level, to assist the user in carrying out complex information seeking activities. Our work is motivated by this earlier work, but we look at information within documents at a more fine grained level for a given search topic.

Recent advances in IR have focused on user centric measures of utility, satisfaction of information for supporting search tasks and information seeking [1, 4, 11]. Several studies have examined a move from document level relevance to within document (sentence and paragraph) level relevance to extract parts of the documents that can satisfy a user's information need [5, 9]. Looking for useful, important and relevant information within a document that can directly provide information for a user search need can be used for the generation of effective document surrogates or information cards and to provide answers to a user's question [2, 3, 14]. Providing relevant and useful information is also important to address the information gaps in a user's knowledge of a topic and to provide better support for learning and gaining knowledge [13, 16].

Hassan et al. defined complex search tasks as a multi-aspect or a multi-step information need consisting of a set of related tasks [6]. In their work, the authors sought to identify and recommend sub-tasks to users based on their search queries in order to provide support in the overall search process. Similarly there has been work on supporting exploratory search and serendipity in web search to address complex search topics [6, 8, 12, 17]. Most of this prior work aimed to support exploratory and investigative activities by grouping user's queries and search behavior, and retrieving documents using task or session level information from the user. Whereas in our work, we try to study the multi-aspect parts of the information contained within potential documents to be presented to the user to address a complex search task. Identifying which information must be shown to users from a multitude of information from potential relevant documents originating from different sources is a challenging task. In this paper, we study how users interpret and perceive information from within retrieved documents for a given search task. To do this, we analyze the user's rating and reasoning while they categorize textual information within retrieved documents for different search topics on 4-levels of relevance and importance. We focus on looking within a document at a granular level of textual (information) units [9] to understand *why* and *which* parts of the document are more relevant and important than others. We believe that understanding what kind of information better supports and satisfies a user's information need effectively can help in the overall search process involving multiple steps.

2 EXPERIMENTAL INVESTIGATION

In this section, we introduce the design of the user study and dataset used for our experiments.

CHIIR 2017 Workshop on Supporting Complex Search Tasks, Oslo, Norway.

Published on CEUR-WS, Volume 1798, http://ceur-ws.org/Vol-1798/.

2.1 User Study Design

Participants were presented with a series of information needs and a single relevant document for each one. Participants were recruited through the *Prolific* crowdsourcing platform[1].

TASK: Assessing already highlighted information units (AIHIU): Participants assessed already highlighted information units on a scale of [1-4]. The first author of this paper manually identified and highlighted topically related textual units from the documents to be categorized by the users between 4 classes of relevance and importance:

(1) C1: Highly relevant and important
(2) C2: Fairly relevant and important
(3) C3: Slightly relevant and important
(4) C4: Neither relevant nor important

Participants were asked to explicitly outline the reasons for their ratings.

We merged the dimensions of relevance and importance in a single scale, as in our related user study [2], users were asked to find and highlight useful and important parts of the document that satisfies and addresses the given information need. In the analysis of the data and user feedback we found that at times users find it difficult to identify information units which are useful and important, since it gets quite ambiguous while performing annotations within document level unless separate topic specific guidelines are provided as was done by Habernal et al. [5]. It would be worth exploring user's annotation for the above mentioned scale of relevance and importance separately. But we believe that it will be quite complex to perform annotations at a granular level within documents for information at varying scales of *relevance, usefulness, importance* separately, unless the definition of these concepts is properly defined for annotations within the document level. This latter issue is an important area to be explored in future work, but is beyond the scope of this paper.

Our study focuses on the following specific research questions:

RQ-1: How do users rate and perceive information within retrieved documents for different types of topics?

RQ-2: How can we categorise user feedback to provide better support for search topics which are explorative and investigative in nature?

Research Contribution: The main contribution of this paper is the categorization and analysis of user feedback while assessing textual units within retrieved documents. Our study draws some important observations and conclusions regarding *why* some parts of a document are more relevant than others (e.g. they provide more contextual information, they contain background information) and *which* types of information appear to be effective for satisfying end users (e.g showing examples, providing facts).

2.2 Dataset and Study Procedure

We used data from the TREC 2012 session track for our study [8]. We selected three information needs from this dataset and at random one relevant document from the *qrels* for each of the these information needs. Since this is a comprehensive and cognitively intensive task for our participants, we opted to concentrate on detailed descriptive analysis of a small number of documents for this initial study.

Differences in the user's topic familiarity can influence their search behaviour [10], thus to ensure participants are familiar with the topics, we carefully chose the following three simple and generic topics from the TREC data set

– T0: Wedding Traditions: web document shown to users majorly contained factual information

– T1: Smoking Cessation: web document shown to users majorly contained recommendation related information

– T2: Junk Food Taxes: web document shown to users majorly contained opinionated and factual information

Table 1: Participants Demographics Information

Study type	Users	Age Range	Demographics	Nationality
AIHIU study	7	20-43	4 M & 3 F	5 UK & 2 US

To carry out the user study, topics were organised and always presented in the same order. Classification of highlighted textual units with reasons were collected using the *Prolific* crowdsourcing platform. Table 1 shows the demographics of the participants. All participants were native English speakers. In accordance with standard crowdsourcing practice for this type of work, participants were paid between 8-9 euros on a per hour basis.

3 EXPERIMENT AND RESULTS

In this study we asked the annotators to classify already highlighted textual units on a scale of [1-4] and provide reasons for their ratings as discussed in Section 2.1.

3.1 Results and Analysis

In Table 2, we present the distribution of user's ratings of highlighted textual units across different topics and all topics combined. We also calculate the mean and variance of the user's ratings to indicate the spread and diversity in the ratings of a particular user. In Table 3, we show the generalization of the users' feedback and reasoning while categorizing textual units into one of the 4-levels of relevance and importance. We had a set of 47 textual units in total for all three documents combined together (T0:14, T1:19 and T2:14) which were annotated by 7 users, thus the dataset of user's feedback had 329 statements. A few statements were repetitive in nature and some were just single word entries such as: *clarification*, *example*, *tips*, etc. We analyzed the user's feedback within separate classes of relevance, where we try to generalize and capture the user feedback effectively by grouping statements in terms of finding answers to *why* some parts of documents are more relevant and important, and *which* types of information appears to be effective in satisfying the end users as shown in Table 3.

Data analysis indicates that while assessing textual units, user feedback and reasoning overlaps over closely related levels of relevance. Most of the information that is marked as *highly relevant and important* is considered as self sufficient, users believe that

[1] https://www.prolific.ac/

[2] Our study, Titled: *Identifying Useful and Important Information within Retrieved Documents*, appears at CHIIR 2017 main conference. This work described in this paper is an extension and detailed analysis of the main study.

it provides a complete meaning by itself and contains necessary sources, references or numbers to backup the statements. Units that are marked as *fairly relevant and important* are considered to contain information related to the topic but are often supplementary in nature and need context to properly express their meaning. Units that are marked as *slightly relevant and important* are often considered to be related to one or more aspects of the topic. Users found that these statements lack proper argument and in some cases need supporting claims and references. Units that are marked as *neither relevant nor important* are often considered as incomplete information or having lack of proper reasoning.

3.2 Discussion

Based on the analysis of user's variations in terms of feedback and ratings as shown in Table-2 and Table-3, we speculate that when participants perceive information, it can broadly be categorised into 4 different types:

1) One who contradicts most of the information
2) One who satisfactorily accepts the information
3) One who is more doubtful, and believe that information might be correct, but wish to get the supporting claims
4) One who finds information to be assumptious (made up), and believe information is not factual

We analyzed the user's specific ratings and feedback across 3 topics as indicated in Table-2. We found that *User-1* for topic: "Wedding traditions", satisfactorily accepts most of the textual units as highly relevant and important as the information was more factual in nature, while for topic: "Smoking Cessation" and "Junk food Taxes" the user considered the information to be assumptious and thus was contradicting with the textual units as the documents on these topics were more opinionated and recommendation based in nature. *User-2* satisfactorily accepts information as highly or fairly relevant and important across all three topics. *User-3* satisfactorily accepts information as highly relevant and important for topics: "Smoking cessation" and "Junk food taxes", but for topic: "Wedding tradition'", slightly misinterpreted the task as discussed below and thus rated many units as neither relevant nor important. Users (4, 5, 6, and 7) critically analyzed the information with proper reasons while generally categorizing units as highly, fairly or slightly relevant and important.

Further, analysis of the feedback reveals that sometimes participants misinterpret the task and develop their own interpretation while analyzing the information within retrieved document. For example when they were asked to look at the information regarding *Topic T0: "Wedding traditions that are interesting and different from what they are used to*, and the document that was shown was a factual one based on the *Japanese wedding and tradition* two users seemed to slightly misinterpret the task. We speculate that user-2 wrongly interpreted the task as categorising textual units based on whether the information is *contemporary* or *traditional* in nature, similarly user-5 categorised the information while doing comparative analysis with western wedding traditions and culture.

This user study opens discussions for future explorations, for example when and how to provide information to users: in more detail, in an abstract way, as a gist or summary depending on the complexity of search tasks and types of document been retrieved containing opinionated, recommendation or factual information. We believe the findings of this work will stimulate discussion on: How can we support users by understanding individual differences and way of interactions within documents?

4 CONCLUSIONS AND FUTURE WORK

The results indicate that annotation varies across users and for same users across different topics as shown in Table 2. The way users perceive information varies depending on the source and type of information such as factual, opinionated, recommendation as explored in our study. Analyzing users' reasoning and their feedback provided some interesting insights on *why* some parts of the document are more relevant and important than others and *which* types of information better satisfy the end users as discussed in Table 3.

This is a preliminary investigation and needs further research and exploration to draw effective conclusion from the studies. This work opens question for future exploration:

1) How to group users based on their behaviour patterns in terms of how they perceive information in documents and support information accordingly for complex search tasks?

2) How can we model information support for different types of topics where *type* and *credibility* of information is in question for, e.g. opinionated, factual, recommendation related, information as used in this study?

3) How can results be presented depending on the type of task in terms of satisfying end users by providing information which is factual, topically relevant, diverse and novel?

In future work, we will further explore the topics opened up in this study with larger numbers of participants. Additionally, the results of this work will contribute to our broader objective of creation of richer document surrogates and summaries, and effective presentation of information to users to promote for effective search and engagement, and emerging areas such as improving learning through search.

Acknowledgment: We thank the reviewers for their feedback and comments. This research is supported by Science Foundation Ireland (SFI) as a part of the ADAPT Centre at Dublin City University (Grant No: 12/CE/I2267).

REFERENCES

[1] A. Al-Maskari and M. Sanderson. A review of factors influencing user satisfaction in information retrieval. *JASIST, 2010*, 61(5):859–868.
[2] P. Arora and G. J. F. Jones. Position paper: Promoting user engagement and learning in search tasks by effective document representation. In *Proceedings of SAL workshop, SIGIR 2016*.
[3] M. S. Bernstein, J. Teevan, S. Dumais, D. Liebling, and E. Horvitz. Direct answers for search queries in the long tail. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 237–246, 2012.
[4] M. Cole, J. Liu, N. Belkin, R. Bierig, J. Gwizdka, C. Liu, J. Zhang, and X. Zhang. Usefulness as the criterion for evaluation of interactive information retrieval. *Proc. HCIR*, pages 1–4, 2009.
[5] I. Habernal, M. Sukhareva, F. Raiber, A. Shtok, O. Kurland, H. Ronen, J. Bar-Ilan, and I. Gurevych. New collection announcement: Focused retrieval over the web. In *Proceedings of SIGIR 2016*, pages 701–704.
[6] A. Hassan Awadallah, R. W. White, P. Pantel, S. T. Dumais, and Y.-M. Wang. Supporting complex search tasks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 829–838, 2014.
[7] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of SIGIR 2000*, pages 41–48.
[8] E. Kanoulas, B. Carterette, M. M. Hall, P. D. Clough, and M. Sanderson. Overview of the TREC 2012 session track. In *Proceedings of The Twenty-First Text REtrieval*

	User-1	User-2	User-3	User-4	User-5	User-6	User-7
User's rating classification for three topics (T0, T1 and T2)							
C1 units	18	40	37	4	19	27	8
C2 units	9	7	2	17	21	6	23
C3 units	10	0	1	26	7	11	14
C4 units	10	0	7	0	0	3	2
Mean	2.25	1.15	1.53	2.47	1.74	1.79	2.21
Variance	1.38	0.13	1.18	0.42	0.49	1.01	0.60
User's rating classification for topic T0: Wedding Traditions							
C1 units	8	12	6	3	2	7	4
C2 units	4	2	1	5	7	1	8
C3 units	1	0	0	6	5	5	0
C4 units	1	0	7	0	0	1	2
Mean	1.64	1.143	2.601	2.21	2.21	2	2
Variance	0.80	0.12	2.10	0.60	0.45	1.14	0.86
User's rating for classification topic T1: Smoking Cessation							
C1 units	7	15	17	0	11	9	1
C2 units	3	4	1	11	7	4	11
C3units	4	0	1	8	1	5	7
C4	5	0	0	0	0	1	0
Mean	2.37	1.21	1.16	2.42	1.47	1.90	2.32
Variance	1.50	0.17	0.24	0.24	0.35	0.94	0.32
User's rating classification for topic T2: Junk Food							
C1 units	3	13	14	1	6	11	3
C2 units	2	1	0	1	7	1	4
C3 units	5	0	0	12	1	1	7
C4 units	4	0	0	0	0	1	0
Mean	2.71	1.07	1	2.79	1.64	1.43	2.29
Variance	1.20	0.07	0	0.31	0.37	0.82	0.63

Table 2: User's rating classification and distribution for different topics, where C1: highly relevant and important, C2: fairly relevant and important, C3: slightly relevant and important, C4: neither relevant and important

Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6-9, 2012, 2012.

[9] M. P. Kato, M. Ekstrand-Abueg, V. Pavlu, T. Sakai, T. Yamamoto, and M. Iwata. Overview of the ntcir-11 mobileclick task. In *NTCIR*, 2014.

[10] D. Kelly and C. Cool. The effects of topic familiarity on information search behavior. JCDL '02, pages 74–75. ACM, 2002.

[11] J. Mao, Y. Liu, K. Zhou, J.-Y. Nie, J. Song, M. Zhang, S. Ma, J. Sun, and H. Luo. When does relevance mean usefulness and user satisfaction in web search? In *Proceedings of SIGIR 2016*, pages 463–472.

[12] G. Marchionini. Exploratory search: From finding to understanding. *Commun. ACM*, 49(4):41–46, Apr. 2006.

[13] S. Y. Rieh, K. Collins-Thompson, P. Hansen, and H.-J. Lee. Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science*, 42(1):19–34, 2016.

[14] M. Shokouhi and Q. Guo. From queries to cards: Re-ranking proactive card recommendations based on reactive search history. In *Proceedings of SIGIR 2015*, pages 695–704.

[15] A. Spink, H. Greisdorf, and J. Bateman. From highly relevant to not relevant: examining different regions of relevance. *Information Processing & Management*, 34(5):599–621, 1998.

[16] P. Vakkari. Searching as learning: A systematization based on literature. *Journal of Information Science*, 42(1):7–18, 2016.

[17] R. W. White and R. A. Roth. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009.

C1 information units	**C2 information units**
Highly relevant and important	**Fairly relevant and important**
Facts	Added details
Examples, tips	Supportive material, tips, advices
Provide context	Suggestions
Show attitudes & opinions	Opinions
Identify commonalities & differences	Quite broad not concrete information
Topically relevant	Details about items or aspects missing
Background information	References
Indicate benefits, outcomes	Not applicable to all (suggestions)
Explain & describes process	Evidence or explanation of an aspect
Provide rationale, motivation	Some aspects (location, time) discussed
	Not very detailed information
	Comparisons
	Discusses changes happening
	Information indirectly related to tasks

C3 information units	**C4 information units**
Slightly relevant and important	**Neither relevant nor important**
Non specific details	Repetitive information
Background, not topically related	Facts and flow missing
Information meaningless out of context	Advices
Possible solutions	Mathematical aspects e.g. increment by 25%
Comparative analysis	Reasons missing
Personalized information	Contextual information missing
Context mismatch	What and why's missing
Forecasts & predictions	Source of information missing
Partial information on certain aspects	
Obvious information	

Table 3: Generalization of user's reason and feedback while annotating textual units within retrieved documents