Why Assessing Relevance in Medical IR is Demanding

Bevan Koopman
Australian e-Health Research Centre, CSIRO
Brisbane, Australia
bevan.koopman@csiro.au

Guido Zuccon
Queensland University of Technology
Brisbane, Australia
g.zuccon@qut.edu.au

ABSTRACT

This study investigates if and why assessing relevance of clinical records for a clinical retrieval task is cognitively demanding. Previous research has highlighted the challenges and issues information retrieval systems are faced with when determining the relevance of documents in this domain, e.g., the vocabulary mismatch problem. Determining if this assessment imposes cognitive load on human assessors, and why this is the case, may shed lights on what are the (cognitive) processes that assessors use for determining document relevance (in this domain). High cognitive load may impair the ability of the user to make accurate relevance judgements and hence the design of IR mechanisms may need to take this into account in order to reduce the load.

Categories and Subject Descriptors: H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval
General Terms: Experimentation.

1. INTRODUCTION

The collection of relevance assessments is important for information retrieval (IR) systems evaluation. Relevance is a complex notion: subjective to the person performing the assessment, dependent on contextual factors and often acting on multiple dimensions (i.e., factors like opinion, readability and trustworthiness may influence a relevance judgement) [3]. To the best of our knowledge, however, there has been little or no work that investigates if and why it is cognitively demanding for assessors to judge relevance.

In this paper, we aim to determine: (i) if assessing document relevance is demanding; if so (ii) what are the indicators of a demanding assessment; and (iii) what are the reasons behind an assessment being demanding or not. Toward these aims, we focus on medical IR, and more specifically on the task of finding patients suitable to clinical trials, i.e., the task modelled in the TREC Medical Records Track (MedTrack) [5]. It has been shown that this is, in general, a difficult task for IR systems due to factors like vocabulary and granularity mismatch, conceptual implication, and inferences of similarity [1]. However, no previous work has explored whether this also applies for humans, and whether assessing the relevance of health records for this task is cognitively demanding (indeed, difficult) for expert assessors.

MedIR July 11, 2014, Gold Coast, Australia.
ACM SIGIR.

Given the familiarity that medical experts have with medical documents, one may posit that the task of assessing relevance in these documents is not demanding for experts. On the contrary, our quantitative and qualitative analysis of a relevance assessment exercise, performed by four experts, revealed that assessing relevance in the medical domain is often demanding: assessments required substantial time to be formed, implying a substantial cognitive load on the assessors. Given this result, we explore and validate a number of factors associated to both queries and documents that contribute to the difficulty of the assessment task, revealing why this task is demanding.

2. EXPERIMENTAL DESIGN

We used data gathered from a previous relevance assessment task [1]. In this previous study, four medical professionals were asked to judge clinical documents taken from the TREC MedTrack collection [5]. As we used data from an existing study not explicitly designed to fully answer the research questions of this paper, we are constrained by the data captured in the previous study. Nevertheless, a number of insights into how demanding assessment are can be derived.

The original TREC MedTrack queries were used and a total of 1030 documents were assessed.[1] To collect assessments, the Relevation! judging system was used [2]. Queries were divided between the four assessors with each query being fully judged by only one assessor. Each assessor also completed two control queries to familiarise themselves with the task. As all assessors completed the same control queries, these were used to determine inter-coder agreement. The test queries were divided so that each assessor judged, in total, roughly an equal number of documents. For each document, judges were asked to mark the document as "highly relevant", "somewhat relevant" or "not relevant" with respect to that query (as per TREC MedTrack guidelines). In addition, using Relevation!, assessors could provide a free-text comment regarding their decision. On completion of judging all documents for a query, the assessor was also asked to answer the following questions about the query: 1) "How difficult was this query to judge?". Choices: "Very difficult", "Moderately difficult" or "Easy". 2) "How would you rate the quality of the assessments you have provided for this query?" Choices: "High quality", "Average in quality" or "Poor quality". 3) "Other comments?" Here judges could provide qualitative comments regarding the particular query.

[1] 4 queries were excluded from the original 85 TREC MedTrack queries as no relevance assessments were collected for these.

As Relevation! is a web-based system, the HTTP access log was used to capture the interaction assessors had with the system. This included which queries and documents they viewed, when documents were judged and, importantly, the timestamps for these events. These timestamps were used to extract the amount of time each assessor spent in judging individual documents.[2] The difference in time between two consecutive HTTP POSTs was used as the measure of time it took to judge that document. On manual review, any time periods greater than 2500 seconds (42 minutes) was indicated as a break (e.g., lunch or coffee) and these timings were excluded. Note that qualitative feedback from assessors (e.g., difficulty and quality) were collected at query level, while quantitative statistics such as time to perform a judgement were collected both at query and document level.

A total of 58 hours (14.5 hours per assessor) of judging was required to complete the 942 documents.[3] The average time spent per document was 3.7 minutes. Using the control queries, inter-coder agreement was found to be 0.85, in-line with an inter-coder agreement of 0.8 found by the TREC MedTrack organisers.[4] Control queries also contained documents already judged by TREC assessors; therefore, if the TREC assessor is added as a fifth assessor, then agreement between all five assessors was 0.80.

3. IS ASSESSING RELEVANCE DEMANDING?

To determine if and why assessing relevance is demanding we analysed: (i) qualitative feedbacks given by assessors in relation to the assessment difficulty of each query; and (ii) the amount of time required to judge documents.

3.1 Did assessors find judging difficult?

Assessors rated each query according to how difficult it was to judge and further provided a self-assessment of the quality of their judgements. Results are shown in Figure 1. Assessors stated that about half of the queries were easy to assess, with the remaining half being of moderate difficulty. Only one query was considered very difficult to judge.[5] Nevertheless, the assessors believed the judgements they provided were of average or high quality. (No queries were marked as low quality.)

While these qualitative assessments are ultimately subjective (the self-perception of difficulty and quality may vary between assessors), it is clear how a significant number of assessments was perceived to be more demanding than others.

3.2 Time as indicator of demand

Beside examining the qualitative feedback of the difficulty in assessing documents, we also consider time as an indicator of judging demand. The intuition is that documents that required more time for assessment are more demanding; similarly, the longer it took on average to judge documents for a query, the more demanding that query.

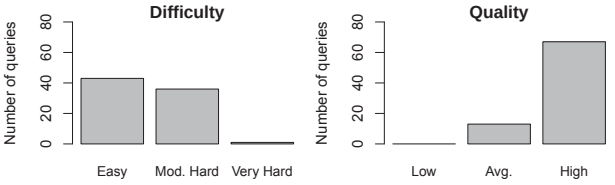

Figure 1: Judges' qualitative feedback on difficulty and quality of their assessments.

Difficulty	#Queries	Median sec./doc
Easy	44	130sec
Moderately Difficult	36	207sec (+59%)
Very Hard	1	219sec (+68%)

Table 1: Timing results by difficulty.

The use of time as an indicator of assessment demand is confirmed by the results of Table 1 that shows the judges' qualitative feedback about query difficulty along with the median document judging time for each difficulty level. This analysis shows that queries judged as moderately difficult took 59% longer to judge than those marked easy, endorsing the intuition that time is a (fine grain) indicator of assessment demand.

4. WHAT INDUCES COGNITIVE LOAD?

4.1 Are longer documents harder to judge?

Smucker & Clarke found that in web search, judging time was mainly influenced by document length [4]. Document length was, therefore, used as the main indicator for their time-biased evaluation measure [4].

In our study, if document length was also a measure of demand, then the Easy/Mod/Hard label assigned by assessors would simply relate to short, moderate and long documents respectively. By extension, shorter documents would be less demanding to judge. However, this was not found to be the case: there was no correlation between time to judge a document and the length of the document ($p = -0.0132$).

4.2 Are documents with discharge summaries easier to judge?

Many of the clinical documents used in our collection contained a discharge summary section.[6] Assessors commented that they often skimmed the document looking for a discharge summary section to read first rather than reading the document from top to bottom. Sometimes the relevance of a document could be determined from reading the discharge summary alone.[7] Based on these comments, we formed the hypothesis that documents containing a discharge summary would be quicker and less demanding to judge. However, our results show the contrary: the median time to judge a document with a discharge summary was 184 sec., vs. 118 sec. for documents without a discharge summary.

[2]The HTTP log is available online at:
https://github.com/ielab/MedIR2014-RelanceAssessment
[3]This number excludes documents from the control queries and those which took more than 2500 seconds to judge (i.e., where the assessors was deemed to have taken a break).
[4]Based on personal communication with Bill Hersh, TREC MedTrack organiser, 29 May 2013.
[5]Query 149: "Patients with delirium hypertension and tachycardia".
[6]A discharge summary is a narrative produced when a patient is discharged from hospital. Discharge summaries provide an overview of the patient's entire stay in hospital.
[7]Note that not all documents contained discharge summaries.

Documents	Time to judge (seconds)			
	mean	stddev	max	min
non-relevant	219	191	1614	5
relevant	224	221	2092	26
highly-relevant	167	209	2092	26
somewhat-relevant	289	217	1314	60

Table 2: Timing results by relevance grade.

4.3 Is the grade of relevance related to cognitive load?

Does the relevance grade of a document (i.e. highly relevant, relevant, not relevant) affect how demanding it is to judge? Table 2 shows the time it takes to judge documents according to the relevance grade. When considering only binary relevance (i.e., relevant vs. non-relevant), the average time to judge relevant and non-relevant documents does not differ significantly, although the time to judge relevant documents varies more (stddev); both the maximum and minimum judging time are greater for relevant documents. In contrast, when graded relevance is considered, some important differences are revealed: highly relevant documents are the *least* demanding to judge, whereas somewhat-relevant documents are the *most* demanding to judge. This finding suggests that clear cases of relevance (highly relevant or non-relevant) are less demanding. What is demanding is judging documents where relevance is less certain: cases where relevance is subjective or where the evidence for relevance is implicit and needs to be inferred. We explore more of these situations in the following section by analysing the assessors qualitative feedback.

5. WHY IS ASSESSMENT DEMANDING?

On completion of judging a query, assessors could optionally provide free-text, qualitative comments regarding their judging of the particular query. Assessors provided these comments for 57 out of 81 (70%) queries. We analysed their comments to gain a greater insight into their rational for assessment and to determine why it might be demanding. Table 3 contains a selection of assessor's comments which will be referred to throughout this section. The assessors' comments were used to identify queries exhibiting the following characteristics: (i) "objective", where the indicator of relevance was clear and explicit according to the assessor; (ii) "temporal", where relevance was strongly dependent on temporal aspects (of query or documents); (iii) "interpretation", where the interpretation of the query was subjective and the assessor had to decide on a particular interpretation; and (iv) "dependent aspects", where there were two or more conditions specified in the query — often dependent on each other — that had to be met. Queries not exhibiting any of the aforementioned characteristics were characteristed as "none". Note, these characteristics were derived from the relevance criteria, as stated in the assessor's comments, and not according to the query keywords. Queries were grouped according to these characteristics and we analysed the average time to judge the documents for queries with that characteristic. This is done to understand if some characteristics — and therefore some queries — were more demanding than others. The average time to judge according to each characteristic is shown in Figure 2.

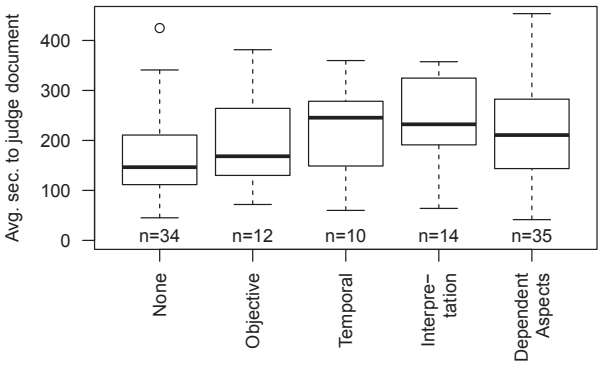

Figure 2: Average time to judge the documents for queries with different characteristics. Queries requiring some "interpretation" on the part of assessors were the most demanding.

Those queries identified as "none" (n=34, 60%) required, on average, the least assessment time and were the least demanding. Queries identified as "objective" (n=12, 21%) were marginally more demanding, as the assessor had a clear criteria to identify relevance and all that was required was to assert if that criteria applied to the particular document.

5.1 The effect of temporality on relevance

For "temporal" queries (n=10, 18%), the assessors specifically cited temporality as an important factor in determining relevance. The most common situation was when information pertaining to the query was found in the patient's past medical history section. Assessors had to decide whether the information was still valid: some conditions are ongoing (e.g., query `162`, Table 3), while others are temporal and are unlikely to still be valid (e.g., query `127`). In certain cases, assessors consulted the actual dates of the past medical history information to determine how recent the information was and whether it might still apply. In other cases, the query was interpreted according to a temporal definition (e.g., query `111`, where the assessor defined 'chronic back pain' as a condition persisting for at least 3 months). Queries exhibiting temporality tended to be the most demanding as assessors had to locate and reason with dates found in the documents.

5.2 Judging was highly subjective

For "interpretation" queries, assessors, at times, discussed their decisions regarding relevance. Although confident in their assessments, they stated that the interpretation of the query was subjective and often required careful consideration regarding different possible interpretations. For example, for query `101`, assessors debated whether a patient born deaf could be considered as exhibiting hearing loss. (Technically, if they never had any hearing, then they never had a loss of hearing.) One assessor thought such a document was relevant, while another assessor thought the document was not relevant. A medical encyclopaedia was consulted and the assessor decided to include patients born deaf as relevant. Queries requiring subjective interpretation showed a higher level of demand compared to other queries.

The task description given to assessors (recruitment of patients matching a certain inclusion criteria for clinical trials [5]) also affected their decisions regarding relevance. Certain documents described patients who had hearing loss

Query		Assessors' Comment
101	Patients with hearing loss	*It was not clear whether you wanted someone with current hearing loss or someone who had experienced reversible hearing loss due to an infection.*
102	Patients with complicated GERD who receive endoscopy	*Complicated GERD is a rather ambiguous term - could use clarification to yield better results (ex. stage a/b/c). Endoscopy is a blanket term for visualisation of a hollow organ - therefore some search results included patients who have had colonoscopies, but not upper endoscopies relevant to GERD.*
103	Hospitalized patients treated for methicillin resistant Staphylococcus aureus MRSA endocarditis	*Treatment of MRSA is the same no matter where it is in the body. Could have picked up a lot of documents because of the treatment regime or MRSA.*
111	Patients with chronic back pain who receive an intraspinal pain medicine pump	*The definition of chronic back pain used for these judgements was "greater than 3 months"*
127	Patients admitted with morbid obesity and secondary diseases of diabetes and or hypertension	*Without dates, it was difficult to ascertain whether or not hypertension and diabetes were secondary to patients' obesity, as is suggested by the query.*
162	Patients with hypertension on antihypertensive medication	*Once diagnosed with hypertension, you are generally considered to have it for the rest of your life …*
171	Patients with thyrotoxicosis treated with beta blockers	*A lot of hits for beta blockers and very few for any thyroid dysfunction.*
182	Patients with Ischemic Vascular Disease	*Straightforward to look at past medical history for coronary artery disease, bypass grafts or stents.*

Table 3: Assessors' qualitative comments regarding their experience judging the particular query.

on admission but the hearing loss was treated and resolved by discharge. In this case, assessors decided these patients would not be eligible for the clinical trial and, therefore, not relevant to the query. For other tasks (for example, finding how hearing loss is treated) these documents may have been highly relevant. These cases highlight the complex and often subjective nature of information need in this domain and that there are often implicit factors in the information need that do not transpire in the query. This further adds to the demand of relevance assessment for these types of queries.

5.3 Queries with dependent aspects

Queries with multiple "dependent aspects" received more debate by assessors and were also among the most demanding and those with the highest variance in judging time. The high variance in time to judge a document is due to the fact that queries with dependent aspects were either: (i) simple to judge, because the assessor just had to ascertain that a document met all aspects; or (ii) demanding to judge, because the assessor had to determine the interaction between the required aspects. Query 171 is an example of the former, simple case. Query 102 is an example of the latter case: GERD[8] is a common condition and is therefore found in many patients' records. The difficulty in interpreting this query was whether the endoscopy was performed because of the GERD or for some other, unrelated condition. There were a number of documents where patients had GERD but received the endoscopy for another reason; these were marked as not relevant. A similar query was 103, where endocarditis and MRSA were mentioned in the same document, but the cause of the endocarditis was not the MRSA. Again, these documents were marked as not relevant. These queries all have multiple dependent aspects to the query; even if both aspects are present in a document, that document may still not be relevant unless the dependence between them can be determined. Determining the dependence often required the assessors to exhaustively search through the document to identify the relationships between the dependent aspects. Doing so required longer judging times and was, therefore, more demanding.

[8]Gastroesophageal reflux disease (GERD) is caused when stomach acid comes up from the stomach into the esophagus.

6. CONCLUSION

Assessing relevance in medical IR is sometimes cognitively demanding and that demand differs depending on queries. Contrary to intuition and previous studies in other domains [4], this study found that document length does not influence demand. On the other hand, the grade of relevance is related with cognitive load (somewhat relevant documents were the most demanding to judge). Characteristics of queries that did increase demand included: temporality, subjectiveness of interpretation and the presence of multiple dependent aspects in the query.

A by-product of this study on what makes a relevance decision demanding, is the identification of some of the aspects that influence a relevance decision (for example, the role of temporality). Future work would, therefore, consider the actual features of the document (for example, temporal ranges or chronic vs. acute conditions) that identify these different aspects affecting relevance.

Data used in this study, including the HTTP interaction log, assessors' comments and qrels, is provided at: http://github.com/ielab/MedIR2014-RelanceAssessment.

Acknowledgements. The authors are grateful to Peter Bruza for his continued mentorship. The relevance assessments were conducted by Timothy Sladden, Warren Brown, Digvijay Khangarot and Thomas Souchen, from the University of Queensland.

7. REFERENCES

[1] Bevan Koopman. *Semantic Search as Inference: Applications in Health Informatics*. PhD thesis, Queensland University of Technology, 2014.

[2] B. Koopman and G. Zuccon. Relevation!: An open source system for information retrieval relevance assessment. In *SIGIR Demo*, Gold Coast, Australia, July 2014.

[3] S. Mizzaro. Relevance: The whole history. *JASIST*, 48(9):810–832, 1997.

[4] M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. In *Proc. of SIGIR*, pages 95–104, Portland, U.S.A, 2012.

[5] E. M. Voorhees and W. Hersh. Overview of the trec 2012 medical records track. In *Proc. of TREC*, 2012.