User Perspectives on Query Difficulty

Lioma, Christina; Larsen, Birger; Schütze, Hinrich

Published in:
Advances in Information Retrieval Theory

DOI:
10.1007/978-3-642-23318-0_3

Publication date:
2011

Document version
Early version, also known as pre-print

User Perspectives on Query Difficulty

Christina Lioma[1], Birger Larsen[2], and Hinrich Schutze[1]

[1] Informatics, Stuttgart University, Stuttgart, Germany
[2] Royal School of Library and Information Science, Copenhagen, Denmark
liomaca@ims.uni-stuttgart.de, blar@iva.dk, hs999@ifnlp.org

Abstract. The difficulty of a user query can affect the performance of Information Retrieval (IR) systems. What makes a query difficult and how one may predict this is an active research area, focusing mainly on factors relating to the retrieval algorithm, to the properties of the retrieval data, or to statistical and linguistic features of the queries that may render them difficult. This work addresses query difficulty from a different angle, namely the users' own perspectives on query difficulty. Two research questions are asked: (1) Are users aware that the query they submit to an IR system may be difficult for the system to address? (2) Are users aware of specific features in their query (e.g., domain-specificity, vagueness) that may render their query difficult for an IR system to address? A study of 420 queries from a Web search engine query log that are pre-categorised as `easy, medium, hard` by TREC based on system performance, reveals an interesting finding: users do not seem to reliably assess which query might be difficult; however, their assessments of which query features might render queries difficult are notably more accurate. Following this, a formal approach is presented for synthesising the user-assessed causes of query difficulty through opinion fusion into an overall assessment of query difficulty. The resulting assessments of query difficulty are found to agree notably more to the TREC categories than the direct user assessments.

Keywords: query difficulty, crowdsourcing, subjective logic

1 Introduction

Information Retrieval (IR) systems aim to retrieve relevant information from a usually large and heterogeneous data repository such as the Web, in response to a user query. Whereas most IR systems aim to employ globally optimal algorithms that can reliably retrieve documents for most queries, there exist some particularly hard queries for which IR systems tend to underperform. Identifying this type of hard queries is important because it allows IR systems to address them in improved ways, for instance by suggesting automatically alternative or additional search terms to the users so that they can reformulate their queries, by expanding the retrieval collection of documents to better answer poorly covered queries, or by training models that can predict further difficult queries [2].

Identifying query difficulty has received a lot of attention (overviewed in Section 2), mainly focusing on factors relating to the system or algorithms used for retrieval, to the properties of the data to be retrieved, or to statistical and/or linguistic features of the queries that make them difficult. This work addresses query difficulty from a different angle, namely the user's own perspectives on query difficulty. Specifically, the research questions investigated are:

1. Are users aware that the query they submit to an IR system may be difficult for the system to address?
2. Are users aware of specific features in their query (e.g., domain-specificity, vagueness) that may render their query difficult for an IR system to address?

The motivation for studying user perspectives on query difficulty partly stems from the fact that increasingly more users regularly use Web IR systems for professional, personal, administrative and further reasons, hence they acquire experience in using search engines. This study investigates whether this search experience can allow users to estimate system-based query difficulty. In addition, the way in which users perceive query difficulty is an interesting question, especially if the users' perspections are found to divert from the system-based understanding of query difficulty, because it can be used constructively in several areas: for instance, when designing user-system interaction functionalities, such as selective user feedback, or when interpreting logged user search sessions and using them to create or train models that involve the user in the search process.

Motivated by the above, this work presents a study using 420 queries from the 2009 TREC Million query track [4], which have already been classified as `easy, medium, hard` by the track's organisers, based on the participating systems performance. A total of 370 anonymous experienced Web search users were recruited through crowdsourcing and asked for their perspectives on the difficulty of these 420 queries. Specifically, users were asked to assess how difficult each query may be for a search engine, without inspecting retrieval results, but simply according to their personal experience and subjective assessment. Furthermore, users were asked to assess, again based on their personal experience and without inspecting retrieval results, whether any of the following causes may render the query difficult for a search engine: the query being too vague, too short, too ambiguous, domain-specific, too specific, or containg typographic errors. Two findings emerge. Firstly, the user-based assessments of query difficulty disagree strongly with the TREC categorisation. Considering the TREC categories as ground truth indicates that users tend to largely underestimate the difficulty of a query for a search engine. Secondly, the user assessments of the causes that may render a query difficult for a search engine are notably more accurate than their overall assessments of query difficulty. In other words, even though users do not seem to reliably assess which query might be difficult, they can assess more reliably which query features might render the query difficult. Following this observation, a formal approach is presented for synthesising the user-assessed causes of query difficulty into an overall assessment of query difficulty. Using probabilistic logic from the subjective logic framework, the individual user-assessed causes of query difficulty are represented as formal

beliefs of query difficulty, which are then fused to produce an expectation that the query is overall difficult. The resulting assessments of query difficulty are found to agree notably more to the TREC categories than the user assessments.

This work contributes an alternative insight into how users perceive query difficulty, which has not been studied before to the best of our knowledge. A formal combination of user perspectives about the causes of query difficulty is presented and juxtaposed to system-based assessments of query difficulty.

The remainder of this paper is organised as follows: Section 2 overviews related work on query difficulty. Section 3 presents the adopted methodology for crowdsourcing user perspectives on query difficulty, and their comparison against TREC categories of query difficulty. Section 4 formalises the user perspectives to induce a probabilistic expectation of query difficulty, which is evaluated against the TREC categories of query difficulty. Section 5 summarises this work and suggests future research directions.

2 Related Work

The study of query difficulty is an active research area in IR, with several applications, such as improving the system's interaction with their users through recommending better terms for query refinement when faced with hard queries [10], providing users with an estimation on the expected quality of results retrieved for their queries, so that they can optionally rephrase difficult queries or resubmit them to alternative search resources, or selectively employing alternative retrieval strategies for particularly hard queries which might be too computationally costly if applied to all queries [2].

Studies of query difficulty can be generally separated into pre-retrieval and post-retrieval approaches (useful overviews are provided in [2,6]). Pre-retrieval approaches focus on features of the query that may render it difficult prior to retrieval, for instance naive features such as query length [14], or indexed statistics of the query terms (e.g., occurrence distribution over the documents in the collection [7], or query term co-occurrence statistics [16]). Further query features include linguistic aspects that may point to difficult queries (e.g. morpheme count per term, count of conjunctions/proper nouns/acronyms/numeral values/unknown words per query, syntactic depth, or polysemy value [11,12]).

Post-retrieval approaches focus on the observed retrieval performance to measure the coherence or clarity of the retrieved documents and their separability from the whole collection of documents [5], or the robustness of the set of retrieved documents under different types of pertrubations [15], or the retrieval status value distribution of the retrieved documents. Furthermore, there exist approaches that combine both pre-retrieval and post-retrieval aspects, for instance the model of Carmel et al., which posits that query difficulty strongly depends on the distances between the textual expression of the query, the set of documents relevant to the query, and the entire collection of documents [3].

Overall, the consensus seems to be that pre-retrieval approaches to query difficulty are inferior to post-retrieval approaches (particularly so when using

linguistic features [12]). A reason for this may be that most queries are very short and hence very poor in features that could potentially discriminate reliably between hard and easy queries. However, pre-retrieval approaches are not as computationally costly as post-retrieval methods, because they do not require dynamic computation at search time.

This work can be seen as a pre-retrieval approach. Its departure from other pre-retrieval approaches is that it does not aim to propose a new improved feature for identifying query difficulty; instead, the aim is to study whether and to what extent users perceive query difficulty. Hence, this work does not use automatic processing to derive features of query difficulty; instead, a large sample of users are asked directly for their opinions regarding whether a query is difficult and which causes might render it difficult. The resulting user perspectives can be potentially useful, both on a theoretical level, for instance to better understand the user's cognitive process during information seeking, and also on a practical level, for instance to improve user-system interaction design functionalities.

3 Crowdsourcing user perspectives

The query set used in this work consists of the 420 queries categorised as `easy, medium, hard` by the 2009 TREC Million Query track [4] organisers, according to the average precision performance of the participating approaches. The distribution of query difficulty in this TREC categorisation is: 29.8% `easy`, 32.1% `medium`, 38.1% `hard` (see Figure 1(a) for the raw counts). These queries have been drawn from a large Web search engine log, without any manual refinement or error correction apart from case collapsing, as described in [1]. For the purposes of this study, user perspectives on the difficulty of these queries were obtained using the Amazon Mechanical Turk (AMT[3]) crowdsourcing platform. AMT is increasingly used to capture and study user preferences or insights into various facets of IR, such as evaluation measures [13]. In this study, 370 experienced Web search engine users were engaged through AMT to:

1. assess the difficulty of a query for a Web search engine, without inspecting retrieval results, but solely according to their personal experience and subjective assessments;
2. assess whether the difficulty of a query may be due to the causes shown in Figure 1(b) or to any other cause that they specify.

The assessments of query difficulty were given in the scale: `easy, medium, hard`, so that they could be directly comparable to the TREC categories. The user assessments of the individual causes that may render a query difficult were binary: `yes,no`. Each query was assessed by 5 users (who had at least $\geq 95\%$ AMT approval rate), resulting in a total of 2100 assessments. The final decision on each query was the most popular among its 5 assessments; in case of draw, another user assessed the query again. Regarding the user statistics, the average user was

[3] https://www.mturk.com

31.7 years old and searched the Web 24.2 days per month on average. 51.5% of the users were native English speakers.

Even though the users were asked to assess query difficulty without inspecting retrieval results, there is no guarantee that they did not do so. A pointer to this direction may be the time they spent on each assessment, which was overall quite low (69.5 seconds on average), leaving little time for inspecting retrieval results.

Finally, an explicit assumption of this study is that query difficulty can be perceived by a user for a query that is not his or her own. For 80.10% of the assessed queries, the participating users explicitly stated that they understood the queries they assessed. Even though understanding a query is not synonymous to cognitively formulating an information need and expressing it as a query, this study uses the former to approximate the latter.

Figure 1(a) shows the categories of query difficulty according to TREC (system-based) versus AMT (user-based). It emerges that users assessed as `easy` more than double the queries categorised as `easy` according to TREC. Furthermore, users assessed as `hard` almost one quarter of the queries categorised as `hard` by TREC. The % of agreement between AMT and TREC is overall low (approx. 34%) and particularly low for hard queries (5%). If the TREC categories are accepted as ground truth, Figure 1(a) seems to indicate that users cannot reliably assess query difficulty, and specifically that they tend to grossly underestimate query difficulty.

Figure 1(b) shows the number (#) and % of queries for which the users identified the causes listed in column 1 as reasons for query difficulty. The three most common causes, sorted decreasingly by frequency, are the query being too vague, too short, and ambiguous. Despite identifying these causes of query difficulty in a query, users did not necessarily assess that query as difficult. This can be seen in Table 1 by comparing the distribution of the queries identified as too vague, too short and ambiguous in the TREC versus AMT categories: the number of vague/short/ambiguous queries increases steadily as one observes the `easy` versus `medium` versus `hard` queries categorised by TREC; however, this is not the case for the AMT assessments, where the number of vague/short/ambiguous queries is the smallest for the `hard` queries, compared to `medium` and `easy` queries. This observation also holds for the other causes of query difficulty. This may be due to the users' poor perception of the (well-known in IR) approximately inverse relation between term occurrence and term discriminativess [8]; users may be more likely to consider easy a term that they are very familiar with through frequent use, than a more discriminative term, and this may affect their estimation about the difficulty of the query containing the term.

The last three columns of Table 1 show the distribution of queries according to the causes of query difficulty only for the subset of queries where TREC and AMT agree. Query vagueness, short length and ambiguity are also the most common causes of difficulty for this subset of queries.

The above observations seem to point to the following paradox: assuming TREC categories as ground truth, user assessments of query difficulty are not accurate; however, user assessments of individual causes that may render queries

difficult are not necessarily inaccurate. This begs the question: can the causes of query difficulty identified by the users be accurately synthesised into an overall estimation of query difficulty? The next section addresses this question.

		AMT							
		easy		medium		hard		$\sum$	
		#	%	#	%	#	%	#	%
TREC	easy	**84**	**20.0**	33	7.9	8	1.9	125	29.8
	medium	84	20.0	**38**	**9.1**	13	3.1	135	32.1
	hard	98	23.3	41	9.8	**21**	**5.0**	160	38.1
	$\sum$	266	63.3	112	26.7	42	10.0	420	100

(a)

Cause	#	%
too vague	107	25.5
too short	84	20.0
ambiguous	69	16.4
domain-specific	45	10.7
has typos	27	6.4
too specific	23	5.5
none	65	15.5

(b)

Fig. 1. (a): Query difficulty according to AMT assessments (based on user perspectives) & TREC categories (based on system performance). Bold font indicates agreement. # indicates number of queries. (b): Reasons for query difficulty based on user perspectives.

	TREC			AMT			TREC & AMT		
cause	easy	medium	hard	easy	medium	hard	easy	medium	hard
	#	#	#	#	#	#	#	#	#
too vague	20	31	56	39	45	23	5	12	13
too short	21	25	38	30	32	22	9	9	11
ambiguous	16	25	28	21	29	19	4	9	7
domain-specific	10	14	21	18	19	8	4	5	5
has typos	5	7	15	9	10	8	3	2	4
too specific	7	6	10	13	7	3	4	1	0

Table 1. Causes of query difficulty for different query groups according to TREC (based on system performance), AMT (based on user perspectives), and the agreement between TREC and AMT.

4 Query difficulty estimation as opinion fusion

This section presents (i) how the subjective perceptions of the users about causes of query difficulty can be formally represented as subjective beliefs (section 4.1); (ii) how the resulting formal beliefs can be fused to give an overall estimation of query difficulty (section 4.2); and (iii) how the resulting formally derived estimation of query difficulty compares to the TREC system-based categorisation of query difficulty (section 4.3).

4.1 Turning user perspectives into formal opinions

Each assessment of the AMT users described in section 3 can be considered as a subjective belief of the user. Using the formalism of subjective logic [9], a frame of discernment can be defined over the proposition that the query is difficult, following [11]. Under this analogy, each of the causes of query difficulty listed in Figure 1(b) can be represented as a different observer holding an opinion about the truth of the proposition that the query is difficult. Subjective logic considers an observer's opinion as decomposable into degrees of belief, uncertainty, and an *a priori* probability in the absence of committed belief mass. These components can be computed directly from the AMT user assessments, using the subjective logic bijective mapping between formal opinion components and observed evidence, defined for binary events [9]. Specifically, the observed evidence can be represented by the `yes, no` assessments of the AMT users described in section 3, denoted Y, N. The belief b and uncertainty u of an opinion can then be estimated as: $b = \frac{Y}{Y+N+2}$ and $u = \frac{2}{Y+N+2}$ (see [9] for a full derivation and explanation of these equations). Hence, the user-assessed causes of query difficulty can be mapped into formal subjective opinions about the query difficulty.

4.2 Fusing opinions of query difficulty using Bayesian consensus

The next step consists in combining the resulting subjective opinions to estimate an overall expectation that the query is difficult. One way of combining these opinions is to assume that they have been formulated independently of each other, that their combination should be commutative, associative and unbiased, and that the uncertainty of at least one of the combined opinions is not zero (because if all opinions have zero uncertainty, they are dogmatic, hence there is no basis for their consensus). Indeed, in this work, the uncertainty of each opinion is uniform and never zero ($u = \frac{2}{7}$ because each query is always assessed by 5 assessors). Then, assumming that A and B represent two different causes of query difficulty, the Bayesian consensus of observers A and B is denoted $\omega^{A,B} = \omega^A \oplus \omega^b$, and its components can be estimated as follows [9]:

$$b^{A,B} = \frac{b^A u^B + b^B u^A}{\kappa} \tag{1}$$

$$u^{A,B} = \frac{u^A u^B}{\kappa} \tag{2}$$

$$a^{A,B} = \frac{a^B u^A + a^A u^B - (a^A + a^B) u^A u^B}{u^A + u^B - 2u^A u^B} \tag{3}$$

where b, d, a denote respectively belief, disbelief, and the a priori probability in the absence of assigned belief mass, and where $\kappa = u^A + u^B - u^A u^B$ ($\kappa \neq 0$). In this work, the *a priori* probability has been set to $a = 0.5$ following [11], so that it is split equally between the two possible states of the frame of discernment, namely that the query either is or is not difficult. The final expectation in the truth of the proposition that the query is difficult is given by:

$$E^{A,B} = b^{A,B} + a^{A,B} u^{A,B} \tag{4}$$

The estimation of query difficulty resulting from Equation 4 is a probability. In order to compare this estimation to the TREC categories of query difficulty, the subjective logic probability needs to be mapped to the `easy, medium, hard` classes of query difficulty. This is done by sorting increasingly all the estimations produced by Equation 4 for all the combinations of causes of query difficulty used in this work, and then binning them into three equal-sized bins. The first, second and third bin respectively contain the lowest, medium, and highest estimations, which are mapped to the `easy, medium` and `hard` classes respectively.

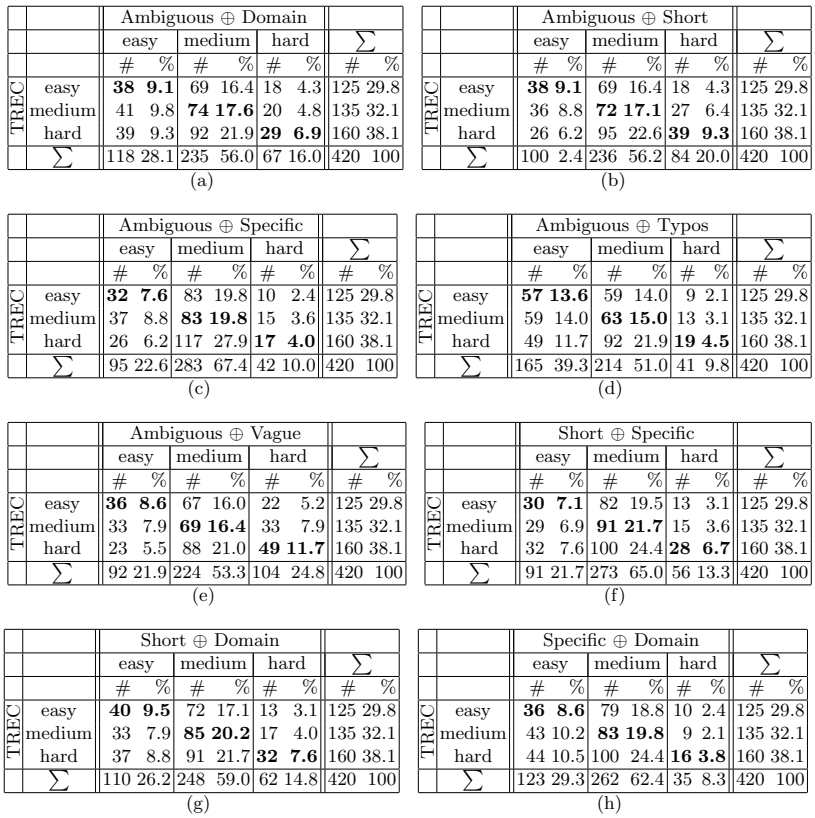
For brevity, combinations of two observers only, which represent pairs of user-assessed causes of query difficulty, are presented in this work. The next section discusses their resulting assessments of query difficulty against the backdrop of the system-based TREC categories.

4.3 Bayesian consensus assessments versus TREC categories

By representing each pair of the six causes of query difficulty listed in Figure 1(b) as observers A and B in Equation 4, 15 Bayesian consensus combinations of pairs of user-assessed causes of query difficulty emerge. Figures 2(a)-3(g) display the categories of query difficulty according to TREC (system-based) versus the assessments of the pairs of causes of query difficulty identified by the AMT users and combined by Bayesian consensus as discussed above. The first row displays the causes of query difficulty that are being combined. The last column is the same for all combinations because it shows the distribution of query difficulty according to TREC (i.e. the ground truth).

Averaging the number of queries assessed `easy` and `hard` for all 15 combinations shown in Figures 2(a)-3(g) reveals that 107.5 queries are now assessed as `easy` using the combinations of causes; this is a notable drop from the direct user assessments which classed 266 queries as `easy` (see Figure 1(a)), and much closer to the number of queries categorised as `easy` by TREC (namely 125). Hence, on average, the subjective logic combinations of user perspectives of query difficulty do not seem to overestimate the number of `easy` queries, like the users themselves did. Furthermore, the average number of queries assessed as `hard` for all 15 combinations is 51.3; this is an increase from the 41 queries that the users directly assessed as `hard`, however it is still much lower than the 160 queries categorised as `hard` by TREC. This indicates that identifying difficult queries is a much harder task than identifying easy queries, when using the combinations of user-assessed perspectives of query difficulty.

The individual combinations of causes of query difficulty are displayed in Figures 2(a)-3(g). Regarding the differences between the individual combinations of causes of query difficulty, Figure 3(h) summarises the number and proportion of queries correctly assessed as `hard` by each of these combinations, using the 160 queries categorised `hard` by TREC as a baseline (see Table 1(a)). The best combination seems to be the user's perception that a query is too short and too vague, which correctly identifies 34.37% of hard queries. Note that the users' direct assessments of query difficulty identified correctly only 13.1% of hard queries. Among the less reliable combinations of query difficulty causes are those



		Ambiguous ⊕ Domain							
		easy		medium		hard		∑	
		#	%	#	%	#	%	#	%
TREC	easy	**38**	**9.1**	69	16.4	18	4.3	125	29.8
	medium	41	9.8	**74**	**17.6**	20	4.8	135	32.1
	hard	39	9.3	92	21.9	**29**	**6.9**	160	38.1
	∑	118	28.1	235	56.0	67	16.0	420	100

(a)

		Ambiguous ⊕ Short							
		easy		medium		hard		∑	
		#	%	#	%	#	%	#	%
TREC	easy	**38**	**9.1**	69	16.4	18	4.3	125	29.8
	medium	36	8.8	**72**	**17.1**	27	6.4	135	32.1
	hard	26	6.2	95	22.6	**39**	**9.3**	160	38.1
	∑	100	2.4	236	56.2	84	20.0	420	100

(b)

		Ambiguous ⊕ Specific							
		easy		medium		hard		∑	
		#	%	#	%	#	%	#	%
TREC	easy	**32**	**7.6**	83	19.8	10	2.4	125	29.8
	medium	37	8.8	**83**	**19.8**	15	3.6	135	32.1
	hard	26	6.2	117	27.9	**17**	**4.0**	160	38.1
	∑	95	22.6	283	67.4	42	10.0	420	100

(c)

		Ambiguous ⊕ Typos							
		easy		medium		hard		∑	
		#	%	#	%	#	%	#	%
TREC	easy	**57**	**13.6**	59	14.0	9	2.1	125	29.8
	medium	59	14.0	**63**	**15.0**	13	3.1	135	32.1
	hard	49	11.7	92	21.9	**19**	**4.5**	160	38.1
	∑	165	39.3	214	51.0	41	9.8	420	100

(d)

		Ambiguous ⊕ Vague							
		easy		medium		hard		∑	
		#	%	#	%	#	%	#	%
TREC	easy	**36**	**8.6**	67	16.0	22	5.2	125	29.8
	medium	33	7.9	**69**	**16.4**	33	7.9	135	32.1
	hard	23	5.5	88	21.0	**49**	**11.7**	160	38.1
	∑	92	21.9	224	53.3	104	24.8	420	100

(e)

		Short ⊕ Specific							
		easy		medium		hard		∑	
		#	%	#	%	#	%	#	%
TREC	easy	**30**	**7.1**	82	19.5	13	3.1	125	29.8
	medium	29	6.9	**91**	**21.7**	15	3.6	135	32.1
	hard	32	7.6	100	24.4	**28**	**6.7**	160	38.1
	∑	91	21.7	273	65.0	56	13.3	420	100

(f)

		Short ⊕ Domain							
		easy		medium		hard		∑	
		#	%	#	%	#	%	#	%
TREC	easy	**40**	**9.5**	72	17.1	13	3.1	125	29.8
	medium	33	7.9	**85**	**20.2**	17	4.0	135	32.1
	hard	37	8.8	91	21.7	**32**	**7.6**	160	38.1
	∑	110	26.2	248	59.0	62	14.8	420	100

(g)

		Specific ⊕ Domain							
		easy		medium		hard		∑	
		#	%	#	%	#	%	#	%
TREC	easy	**36**	**8.6**	79	18.8	10	2.4	125	29.8
	medium	43	10.2	**83**	**19.8**	9	2.1	135	32.1
	hard	44	10.5	100	24.4	**16**	**3.8**	160	38.1
	∑	123	29.3	262	62.4	35	8.3	420	100

(h)

Fig. 2. Query difficulty according to TREC categories (based on system performance) versus query difficulty according to subjective logic predictions based on the following combinations of causes of query difficulty (identified by AMT users): (a): ambiguous & domain-specific; (b): ambiguous & too short; (c): ambiguous & too specific; (d): ambiguous & has typos; (e): ambiguous & too vague; (f): too short & too specific; (g): too short & domain-specific; (h): too specific & domain-specific.

		Typos ⊕ Domain						∑	
		easy		medium		hard			
		#	%	#	%	#	%	#	%
TREC	easy	**57**	**13.6**	64	15.2	4	0.9	125	29.8
	medium	61	14.5	**68**	**16.2**	6	1.4	135	32.1
	hard	69	16.4	77	18.3	**14**	**3.3**	160	38.1
	∑	187	44.5	209	49.8	24	5.7	420	100

(a)

		Vague ⊕ Domain						∑	
		easy		medium		hard			
		#	%	#	%	#	%	#	%
TREC	easy	**35**	**8.3**	72	17.1	18	4.3	125	29.8
	medium	33	7.9	**80**	**19.0**	22	5.2	135	32.1
	hard	23	5.5	96	23.0	**41**	**9.8**	160	38.1
	∑	91	21.7	248	59.0	81	19.3	420	100

(b)

		Vague ⊕ Specific						∑	
		easy		medium		hard			
		#	%	#	%	#	%	#	%
TREC	easy	**25**	**6.0**	84	20.0	16	3.8	125	29.8
	medium	23	5.5	**96**	**23.0**	16	3.8	135	32.1
	hard	14	3.3	120	28.6	**26**	**6.2**	160	38.1
	∑	62	14.8	300	71.4	58	13.8	420	100

(c)

		Vague ⊕ Typos						∑	
		easy		medium		hard			
		#	%	#	%	#	%	#	%
TREC	easy	**47**	**11.2**	69	16.4	9	2.1	125	29.8
	medium	51	12.1	**70**	**16.7**	14	3.3	135	32.1
	hard	28	6.7	107	25.5	**25**	**6.0**	160	38.1
	∑	126	30.0	246	58.6	48	11.4	420	100

(d)

		Short ⊕ Typos						∑	
		easy		medium		hard			
		#	%	#	%	#	%	#	%
TREC	easy	**63**	**15.0**	54	12.9	8	1.9	125	29.8
	medium	57	13.6	**63**	**15.0**	15	3.6	135	32.1
	hard	48	11.4	93	23.0	**19**	**4.5**	160	38.1
	∑	168	40.0	210	50.0	42	10.0	420	100

(e)

		Short ⊕ Vague						∑	
		easy		medium		hard			
		#	%	#	%	#	%	#	%
TREC	easy	**37**	**8.8**	69	16.4	19	4.5	125	29.8
	medium	27	6.4	**73**	**17.4**	35	8.3	135	32.1
	hard	21	5.0	84	20.0	**55**	**13.1**	160	38.1
	∑	85	20.2	226	53.8	109	26.0	420	100

(f)

		Specific ⊕ Typos						∑	
		easy		medium		hard			
		#	%	#	%	#	%	#	%
TREC	easy	**54**	**12.9**	67	16.0	4	0.9	125	29.8
	medium	69	16.4	**61**	**14.5**	5	1.2	135	32.1
	hard	63	15.0	87	20.7	**10**	**2.4**	160	38.1
	∑	186	44.3	215	51.2	19	11.5	420	100

(g)

Queries correctly assessed as hard		
assessment type	#	%
1. user direct assessment	21	13.1
2. formal combinations of causes:		
-too specific ⊕ has typos	10	6.25
-has typos ⊕ domain-specific	14	8.75
-too specific ⊕ domain-specific	16	10.00
-ambiguous ⊕ too specific	17	10.62
-ambiguous ⊕ has typos	19	11.87
-too short ⊕ has typos	19	11.87
-too vague ⊕ has typos	25	15.62
-too vague ⊕ too specific	26	16.25
-too short ⊕ too specific	28	17.50
-ambiguous ⊕ domain-specific	29	18.12
-too short ⊕ domain-specific	32	20.00
-ambiguous ⊕ too short	39	24.37
-too vague ⊕ domain-specific	41	25.62
-ambiguous ⊕ too vague	49	30.62
-too short ⊕ too vague	55	34.37

(h)

Fig. 3. Query difficulty according to TREC categories (based on system performance) versus query difficulty according to subjective logic predictions based on the following combinations of causes of query difficulty (identified by AMT users): (a): has typos & domain-specific; (b): too vague & domain-specific; (c): too vague & too specific; (d): too vague & has typos; (e): too short & has typos; (f): too short & too vague; (g): too specific & has typos. Table (h) displays the number and proportion of queries that have been assessed correctly as hard (using the 160 queries classed hard by TREC as ground truth), firstly by the users when asked directly, and secondly by formally combining the causes of query difficulty perceived by users.

involving the query having typographical errors, being too specific, and being domain-specific. These three causes are also the least frequent in the query set (see Figure 1(b)), being found respectively in only 6.4%, 5.5%, and 10.7% of all queries, which might affect the overall reliability of their combined assessment to a certain extent.

5 Conclusion

This work investigated the users' perceptions of whether a query may be difficult for an IR system to process, and for which causes. 370 anonymised Web search users were recruited using the Amazon Mechanical Turk crowdsourcing platform, and asked to assess the difficulty of 420 Web search queries without inspecting the results retrieved for these queries, but solely according to their subjective opinions and personal experience with search engines. The queries were previously classed as `easy, medium, hard` by TREC as part of the 2009 Million Query track. Considering the TREC categories as ground truth revealed an interesting paradox: when asked to estimate the difficulty of a query, users gave overall inaccurate assessments, largely underestimating hard queries; however, when asked to assess the individual causes that render a query difficult, user assessments largely improved. One plausible reason for this may be the users' incomplete understanding of the (well-known in IR) inverse relation between term occurrence and discriminativeness. In order to investigate further the user-perceived causes of query difficulty, a formal approach was taken, whereby user perceptions were represented as subjective beliefs in the framework of subjective logic. These beliefs were then fused using the Bayesian consensus operator, to produce estimates of overall query difficulty. The resulting estimates were found to be notably better than the direct user assessments, improving the proportion of correctly assessed hard queries from 13.1% up to 34.37%.

The main contribution of this work is in casting light into the user perceptions of query difficulty, and in comparing them to a system-based understanding of query difficulty. Future work includes investigating users' perceptions of query difficulty in relation to their own information needs, to see whether their assessments are more closely related to a system-based understanding of query difficulty, and to find ways of practically applying the user perceptions of query difficulty to improve user-system interaction design for cases of difficult queries. One possible way of doing this is by applying the subjective logic formalism presented here to represent and fuse different aspects of subjective user perceptions.

Acknowledgements This research was partially supported by the Tools for Integrated Search project funded by Denmark's Electronic Research Library (grant number 2007-003292), and by the Relevance of Information Searched in Context project funded by the Research Council of the Danish Ministry of Culture (grant number 2008-001573).

References

1. J. Allan, B. Carterette, B. Dachev, J. A. Aslam, V. Pavlu, and E. Kanoulas. Million query track 2007 overview. In E. M. Voorhees and L. P. Buckland, editors, *TREC*, volume Special Publication 500-274. National Institute of Standards and Technology (NIST), 2007.
2. D. Carmel and E. Yom-Tov. *Estimating the Query Difficulty for Information Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2010.
3. D. Carmel, E. Yom-Tov, A. Darlow, and D. Pelleg. What makes a query difficult? In *SIGIR*, pages 390–397, 2006.
4. B. Carterette, V. Pavlu, H. Fangz, and E. Kanoulas. Overview of the trec 2009 million query track. In E. M. Voorhees and L. P. Buckland, editors, *TREC*, volume Special Publication 500-277. National Institute of Standards and Technology (NIST), 2009.
5. S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIR*, pages 299–306, 2002.
6. C. Hauff. *Predicting the Effectiveness of Queries and Retrieval Systems*. PhD thesis, University of Twente, 2010.
7. B. He and I. Ounis. Inferring query performance using pre-retrieval predictors. In *SPIRE*, pages 43–54, 2004.
8. K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–20, 1972.
9. A. Josang. A logic for uncertain probabilities. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 9(3):279–311, 2001.
10. G. Kumaran and J. Allan. Selective user interaction. In M. J. Silva, A. H. F. Laender, R. A. Baeza-Yates, D. L. McGuinness, B. Olstad, Ø. H. Olsen, and A. O. Falcão, editors, *CIKM*, pages 923–926. ACM, 2007.
11. C. Lioma, R. Blanco, R. M. Palau, and M.-F. Moens. A Belief Model of Query Difficulty that Uses Subjective Logic. In L. Azzopardi, G. Kazai, S. E. Robertson, S. M. Rüger, M. Shokouhi, D. Song, and E. Yilmaz, editors, *ICTIR*, volume 5766 of *Lecture Notes in Computer Science*, pages 92–103. Springer, 2009.
12. J. Mothe and L. Tanguy. Linguistic features to predict query difficulty - a case study on previous TREC campaigns. In *SIGIR Workshop on Predicting Query Difficulty: Methods and Applications*, 2005.
13. M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In F. Crestani, S. Marchand-Maillet, H.-H. Chen, E. N. Efthimiadis, and J. Savoy, editors, *SIGIR*, pages 555–562. ACM, 2010.
14. E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In *SIGIR*, pages 512–519, 2005.
15. Y. Zhou and W. B. Croft. Ranking robustness: a novel framework to predict query performance. In *CIKM*, pages 567–574, 2006.
16. Y. Zhou and W. B. Croft. Query performance prediction in web search environments. In *SIGIR*, pages 543–550, 2007.