Why do you Think this Query is Difficult? A User Study on Human Query Prediction

Stefano Mizzaro
University of Udine
Via delle Scienze, 206
Udine, Italy
mizzaro@uniud.it

Josiane Mothe
Univ. de Toulouse, ESPE, IRIT
UMR5505 CNRS, 118 route de Narbonne
Toulouse, France
josiane.mothe@irit.fr

ABSTRACT

Predicting if a query will be difficult for a system is important to improve retrieval effectiveness by implementing specific processing. There have been several attempts to predict difficulty, both automatically and manually; but without high accuracy at a pre-retrieval stage. In this paper, we focus rather on understanding *why* a query is perceived by humans as difficult. We ran two separated but related experiments in which we asked humans to provide both a query difficulty prediction and reasons to explain their prediction. Results show that: (i) reasons can be categorized into 4 classes; (ii) reasons can be framed into closed questions to be answered on a Likert scale; and (iii) some reasons correlate in a coherent way with the human predicted numerical difficulty. On the basis of these results it is possible to derive hints to be provided to help users when formulating their queries and to avoid them to rely on their wrong perception of difficulty.

Keywords

Information retrieval; Query difficulty; difficulty understanding

1. QUERY DIFFICULTY PREDICTION

One of the outcomes of IR international evaluation campaigns is that system and query variability is high [5]. However, while there is some variability, some queries are difficult or easy for all the participants. For example in TREC Web 2014, which uses the ClueWeb 2012 corpus, the average ERR@20 for topic 278 "What are the lyrics to the theme song for "Mister Rogers' Neighborhood"?" is 0.0048 while the best run for that topic got 0.0820; this is a difficult topic for all systems. On the opposite, for topic 298 "medical care and jehovah's witnesses", the average ERR@20 is 0.5887 and the median is 0.5790; this is an easy topic.

The Reliable Information Access (RIA) workshop [5, 6] has been the first large scale attempt to try to *understand* query (and system) variability and difficulty. The two main conclusions of the failure analysis were: systems were missing an aspect of the query, generally the same aspect for all the systems, and "if a system can realize the problem associated with a given topic, then for well over half the topics studied, current technology should be able to improve results significantly" [6]. When considering failure analysis, 10 classes of topics were identified manually, but no indications were given on how to automatically assign a topic to a category.

Following these findings, there have been many attempts to automatically predict query difficulty. The purpose of a query difficulty predictor is to decide whether a system is able to properly answer the current query [2]. Different kinds of automatic predictors have been proposed in the literature both pre- [7] and post-retrieval [10], based on statistics only or considering some linguistic features [9]. Automatic predictors correlate with actual or observed system effectiveness, but the correlation is always weak, even if it is slightly higher when considering post-retrieval predictors than pre-retrieval ones (although post-retrieval predictors are less interesting, because more costly, than pre-retrieval) [7, 9, 10]. These results limit their practical use in real applications.

Another research direction is addressed by Hauff *et al.* who analyzed the relationship between predictions by non IR expert users and system effectiveness [8]. The authors considered various queries for a single topic or information need and measured the ability of users to judge the quality of query suggestions. They found that: (i) users are not good at predicting system failure; and (ii) the correlations between the users' prediction and both system effectiveness and automatic predictors are weak. We also had similar results when asking annotators to predict query difficulty, under several different experimental conditions, with different user groups, and both from the crowd and from participants in laboratory experiments.

In this paper, rather than focusing on query difficulty rating, we focus on reasons why users think a query is going to be easy or difficult for a search engine. We decided to consider users who are not necessarily IR experts since the latter know how systems work and may not be representative of a large variety of real search engine users. Therefore, when compared to RIA [5, 6], we ask to non experts to provide reasons that explain query difficulty (or ease). When compared to Hauff *et al.* [8], rather than just asking for ratings, we focus on explanations and comments on difficulty.

To know more on these reasons, we performed a user study made up of the two experiments described in the next two sections.

2. EXPERIMENT 1: ELICITING REASONS

The first experiment aimed at eliciting free text reasons why queries are perceived as difficult or easy by users.

2.1 Experimental Design

While we were interested mainly in the reasons why users think a query is going to be easy or difficult for a search engine, the task for human annotators was both to predict the difficulty a search engine may encounter to answer an information need and to explain

SIGIR '16, July 17-21, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2911451.2914696

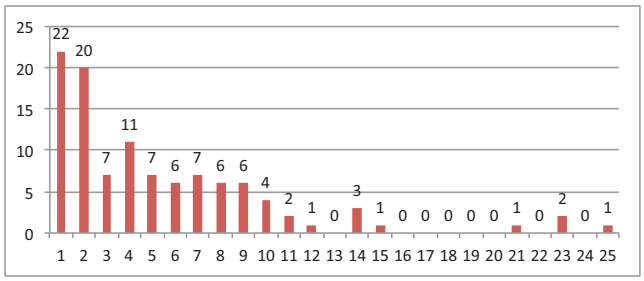

Figure 1: Number of queries by annotation frequency

the reasons of their prediction. Indeed, it is probably more natural for annotators to choose an explicit rating first and then to focus on providing the reasons for it. When asked to annotate the topic difficulty, the main question the annotators had to think of was: is the system going to succeed/fail when processing this query? Does the annotator think the system will retrieve relevant information (an easy query) or not (difficulty query)?

Annotators were provided with the query (TREC topic title); they had to decide the difficulty on, and to comment on the difficulty rating they chose, using the query only. Annotators were asked to use a three level numeric scale: 1 for easy query, 2 for medium query, and 3 for difficult queries. They also could use 0 when they did not know, although they were encouraged to decide on the difficulty. In addition to grading the query difficulty, annotators were asked to indicate the reasons why they thought the query was easy/difficult. For a query and whatever the grade they gave, they could indicate both comments, on its difficult and easy natures. We did not provide any guidance to write the comments apart from using the keyword "Easy:" or "Difficult:" before any comment they write. The tool we provided does not allow them to go back to an annotation they had done previously.

The group of annotators was composed of 38 Master's Students (25 1st and 13 2nd year) in *library and teaching studies*; although they had been trained to use specialized search engines, they had just an introduction class on how search engines work. Annotators could choose as many topics they wanted from a set of 150 TREC topics. Topics were displayed in different order to avoid any bias, as the first topics may be treated differently because the task was new for annotators. Moreover, annotators could skip some topics if they wish; this was done to avoid them to work on a topic they did not understand or felt uncomfortable with. Since the annotation process is difficult, we tried to provide to the annotators the most favorable conditions. The drawback is that the number of annotations varies over topics; this makes numerical analyses more difficult (for example, when computing an average difficulty score the averages are computed over different numbers of scores) but we were here more interested in the reasons, and this kind of qualitative data is less prone to such difficulties.

In this experiment, we used TREC 6, 7, and 8 adhoc task topics.Although these collections are old, we think that the elicitation of reasons will not differ much with other collections. Of course, the annotators knew that the document collection is composed of newspaper articles from the 90s even though it may be difficult for an annotator nowadays to get a picture of what were the popular topics in newspapers more than twenty years ago.

2.2 Results and Analysis

We analyzed the comments that the annotators associated with the evaluation of difficulty. The objectives were (i) to see if there were some recurrent patterns, and (ii) to extract some trends in the comments associated with classes of query difficulty as perceived by users/annotators.

Table 1: Distribution of grades used by annotators

Grade:	Easy	Medium	Difficult	Don't know
Frequency:	227	188	140	17

2.2.1 Descriptive Statistics

Figure 1 shows the number of queries as a function of the number of times it has been annotated by any of the 38 annotators. For example, 22 queries have been annotated a single time (left part of the figure); the most annotated query has been annotated 25 times (right side of the figure). In total 107 queries have been annotated at least by one annotator and 65 three times or more. Table 1 shows the distribution of grades, i.e., the number of times a given grade has been used (e.g., grade Easy has been used 227 times whatever the annotator and the query are).

We collected 460 annotations in total (one annotation count for one topic, one annotator). 107 topics have been graded by at least one annotator; a little fewer have been commented (6 topics have no comment associated with them). It is of course possible that other comments might be generated for the other topics, but with around 70% of the topics being annotated we can be rather confident that most of the reasons have been elicited in this experiment.

2.2.2 Recoding the Free Text Comments

We recoded the free text comments. Table 2 shows some examples of recoding that was made. Each comment could be recoded into more than one recoded phrases; for example the comment "terms are too general, there will be many documents retrieved" has been recoded into *Too-General-Words* and *Many-Documents*. We found out that there were mostly four types of comments, as shown in the table: T (on the topic itself), Q (on the query), W (on the words used), and D (on the documents or on the collection, e.g., if the annotator thinks that some document exists in the collection).

The table also shows how many comments were recoded in each category; however, notice that recoding is always subjective and another recoder may have recoded differently, thus these numbers just provide trends. In a few cases (about 5%), the comment was not explicitly associated with one of these classes. This was for example the case when annotators wrote *vague* without detailing if it was a query term which they found vague or the topic itself. A concrete example is the one of Query 417 (Title: *creativity*) for which the 5 annotators considered the query as difficult using comments such as "too broad, not enough targeted", "far too vague",

Table 2: Examples of recoding, grouped into four categories.

Free text comment	Recoded phrase
T: Topic (274 comments, 35 recoded phrases)	
"The topic is precise"	*Precise-Topic*
Q: Query (142 comments, 23 recoded phrases)	
"The query is formulated in a clear way"	*Clear-Query*
"The query is not precise at all"	*Broad-Query*
W: Words (180 comments, 28 recoded phrases)	
"A single word in the query"	*1-Word*
"The term 'exploration' is polysemous"	*Polysemous-Word*
D: Documents (143 comments, 15 recoded phrases)	
"Risk of getting too many results"	*Too-Many-Documents*
"There are many documents on this"	*Many-Documents*

Table 3: Most frequent comments.

Easy because		Difficult because	
Precise-Topic	66	*Risk-Of-Noise*	50
Many-Documents	45	*Broad-Topic*	43
No-Polysemous-Word	31	*Missing-Context*	34
Precise-Words	25	*Polysemous-Words*	22
Clear-Query	19	*Several-Aspects*	20
Usual-Topic	16	*Missing-Where*	16

"far too vague topic", "keyword used very broad, risk of noise", and "a single search term, risk of getting too many results". While the last comments are directed to one or the other class, the first two are not. We notice that the four categories are roughly equally distributed and have a good coverage.

After recoding, we got 740 annotations for 572 graded queries (each query could be annotated by various annotators; in addition several recoded phrases can be associated with a single comment). For recoding we used 105 different recoding phrases (4 of which are not associated with any of the four categories).

2.2.3 Comments Associated with Ease and Difficulty

Table 3 shows the most frequent recoded phrases associated with ease (left part) and difficulty (right part). Remember that a given query can be annotated by some comments associated with both. For example, while *Precise-Topic* is generally associated with ease (66 times), it is also associated with difficulty in 3 cases. In that case it is associated with other comments, e.g. "The topic is very precise but it may be too specific". In the same way, *Many-Documents* is mostly associated with ease and *Too-Many-Documents* to difficulty, although *Many-Documents* is also associated with difficulty (users may have in mind either a recall-oriented or a precision-oriented task). Also, when *Many-Documents* is used associated with difficulty, it is generally associated with *Risk-Of-Noise*.

2.2.4 Annotator Effect

It may be that some annotators are more likely to use some types of comments than others, either because of what they think about how systems work or because of their search experiences.

We analyzed the link between annotators and the four categories of comments, i.e., associated with Word (W), Query (Q), Topic (T), and Document (D). We grouped the recoded phrases belonging to each category and built a matrix that associates annotators and the four categories of comments associated with. We then used Correspondence Analysis (CA) [1] on that matrix to visualize in one shot the relationships. CA is close to Principal Component Analysis (PCA) in its principle and appropriate when analyzing categorical variables which is the case here. Compared to PCA, CA allows to display in the same space the variables and observations (rows and columns). The distance between objects of any kind is meaningful.

Figure 2 shows the two first axes of the corresponding CA. The horizontal axis divides the figure into two parts. The top part is more related to comments on W and T and so are the annotators in this part of the figure. The bottom part is related to Q and D categories; so are the annotators displayed in this part of the figure. Moreover, the bottom left corner of the figure is more related to comments on D, as the annotators who are in the same corner while the bottom right part is more associated with comments on Q. The annotators near the origin of the axes use annotations from all four categories, while the annotators in the top right corner are more inclined to use comments on Q (according to the horizontal axis) and on W (vertical axis), but do not use comments on D.

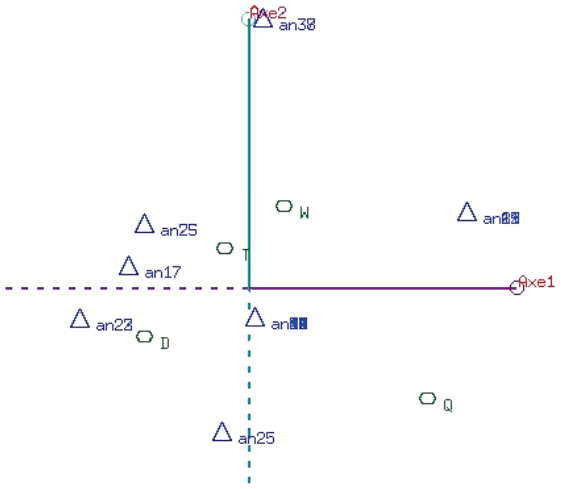

Figure 2: CA, first components: annotators (triangles) and comment categories (ellipses).

We went a step further to check if some annotators use some comments more than others by using the comments rather than the category of comments and found that it is indeed the case too.

3. EXPERIMENT 2: TESTING REASONS

The experiment described in the previous section allowed us to elicit reasons. However, free text questions have two drawbacks: first they need to be recoded and recoding is subjective, and second, annotators are sensitive to different features as we have shown in section 2.2.4. Based on these results, we decided to consider, rather than free text, closed and mandatory questions for the annotators to answer. Closed so that recoding will not be needed anymore, and mandatory so that the annotator effect will be less important. This is done in the second experiment, described in this section.

3.1 From Categories to Questions

We considered each of the 105 recoding phrases obtained in the previous experiment and transformed them into 32 closed questions (denoted with Qi in the following) that could be answered following a Likert scale. Examples of such questions are shown in the first column of Table 4. When recoding, we had tried to keep as most as possible the nuances annotators expressed. When rephrasing into questions, we removed these nuances to limit the number of questions and to remove some redundancy (e.g., precise, specific, focused, delimited, and clear were merged together).

3.2 Experimental Design

We performed a laboratory user study with 22 new volunteers mainly from our research institutes. They were recruited using generic emailing lists and they got a coupon for participating. Each of them was asked to annotate 10 queries (provided in a random order). We used 25 topics from TREC Web 2014 that uses ClueWeb 2012 corpus: we picked up the easiest 10, the most difficult 10, and the medium 5 according to the topic difficulty order presented in the overview paper [3]. We changed from TREC 6, 7, 8 to ClueWeb queries since the latter are more recent, might reflect more the types of current queries on the web, and are now more used in the IR community. Also, in the previous experiment some topics were not annotated: clearly the young students were not at ease with (some of the) old topics.

In order to make the statistical analysis more smooth and sound

Table 4: Examples of questions (column 1) with their Pearson's correlations with human predicted difficulty (col. 2) and actual difficulty (col. 3). Bold indicates a p-value < 0.05, * <0.005.

Question	Correl.	
Q1: The query contains vague word(s)	**.52**	-.30
Q3: The query contains word(s) relevant to the topic/query	**-.41**	.43
Q10: The topic is unusual/uncommon/unknown	**.52**	.26
Q13: The topic has several/many aspects	**.61***	-.07
Q17: The topic is usual/common/known	**.62***	-.25
Q18: The number of documents on the topic in the web is high	**-.69***	-.34
Q19: None or very few relevant documents will be retrieved	**.88***	.32
Q20: Only relevant documents will be retrieved	**-.47**	.09
Q23: Many of the relevant documents will be retrieved	**-.86***	-.20
Q24: Many relevant documents will be retrieved	**-.87***	-.21
Q26: The number of query words is too high	**.62***	**.45**
Q28: The query contains various aspects	**.46**	-.12
Q30: The query is clear	**-.53**	.30

we collected the same number of predictions for each query; we thus consider the same number of annotators for each topic. Annotators had to annotate the level of difficulty of the query, but rather than asking them to provide the reason of their grading in free text only, we asked to answer the 32 predefined questions Qi using a five level scale, from -2 "I strongly disagree" to +2 "I strongly agree". With 32 Qi by 25 topics by 8 annotators each, we collected a total of 6400 reason ratings. The free text reasons they provided has not been analyzed yet.

3.3 Results

Table 4 shows on the second column the Pearson's correlation between the questions and the human prediction of difficulty. Only the 13 Qi having a statistically significant correlation (p-value $<$ 0.05) are included. The seven Qi having values labeled with a * (Q13, Q17, Q18, Q19, Q23, Q24, Q26) have a correlation higher than 0.60 with a p-value $<$ 0.005. These 13, and especially 7, Qi represent the reasons that, according to the users, correlate most with query difficulty. For example, users think a query is difficult because the topic has many aspects (Q13).

However, none of these 13 reasons that users think correlated with query difficulty, turns out to correlate with actual difficulty, with just one exception (Q26). This is shown in the third column in the table that reports the correlation with observed difficulty based on system effectiveness. We used the average ERR@20 as system effectiveness measure, calculated considering all the participant runs. Moreover, the correlation between the human prediction and actual difficulty is low (0.238, p-value 0.25), indeed a much lower value (and not statistically significant as well) than the correlation between Q26 and actual effectiveness, which is 0.45, p-value $<$ 0.05. This means that to obtain a prediction of difficulty, it is much better to ask Q26 than to directly ask for a difficulty rating.

There is also the possibility that some Qi can be combined, maybe also with the difficulty prediction rating, and/or with automatic predictors, to obtain a more accurate numerical prediction. We do not have space here to present those results, but we note (see Figure 3) that some of the seven questions are redundant, and therefore there is no need to require the user to answer all of them.

4. DISCUSSION AND CONCLUSIONS

Other studies have shown that humans are bad query difficulty predictors [8] and generally think queries are easier for systems than they actually are. This paper is a first contribution to try to understand why users (rather than IR experts) think a query is easy or difficult for a search engine. In a first user study, reasons were elicited from free text comments on query difficulty. After a manual recoding and a deep analysis, we found a set of 105 reasons classified into four categories. We then framed 32 closed questions that cover all the mentioned reasons and can be answered through a Likert scale. We used those 32 questions in a second user study, and showed that thirteen correlate with the human prediction of difficulty, and seven of them highly. On the other hand, these questions do not correlate with observed system difficulty. These questions can thus be seen as explanation why humans wrongly think queries are easy or difficult. These results can be useful when training search engine users [4], e.g., to help them to formulate queries and to provide them with hints about their wrong perception of system effectiveness. For example, a user can be told (by a teacher or a search engine) that the fact the query seems usual or common is not linked to system effectiveness.

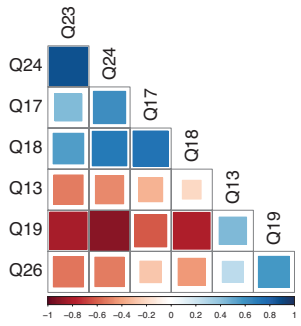

Figure 3: Pearson correlations between questions. Q18, Q19, Q23, and Q24 have a high correlation; they are redundant.

References

[1] J.-P. Benzécri et al. *Correspondence analysis handbook*. Marcel Dekker New York, 1992.

[2] D. Carmel and E. Yom-Tov. *Estimating the query difficulty for information retrieval*. Morgan & Claypool, 2010.

[3] K. Collins-Thompson, C. Macdonald, P. Bennett, F. Diaz, and E. Voorhees. TREC 2014 Web Track Overview. In *Text REtrieval Conference*, 2015.

[4] E. Efthimiadis, J. M. Fernández-Luna, J. F. Huete, and A. MacFarlane. *Teaching and learning in information retrieval*. Vol. 31. Springer Science & Business Media, 2011.

[5] D. Harman and C. Buckley. The NRRC reliable information access (RIA) workshop. In *Conf. on Research and Development in Inf. Retrieval, SIGIR*, pages 528–529. ACM, 2004.

[6] D. Harman and C. Buckley. Overview of the reliable information access workshop. *Information Retrieval*, 12(6):615–641, 2009.

[7] C. Hauff, D. Hiemstra, and F. de Jong. A survey of pre-retrieval query performance predictors. In *Conf. on Information and Knowledge Manag., CIKM*, pages 1419–1420, 2008.

[8] C. Hauff, D. Kelly, and L. Azzopardi. A comparison of user and system query performance predictions. In *Conf. on Inf. and knowledge management, CIKM*, pages 979–988, 2010.

[9] J. Mothe and L. Tanguy. Linguistic features to predict query difficulty. In *Predicting query difficulty Wp, Conf. on Research and Development in IR, SIGIR*, pages 7–10, 2005.

[10] A. Shtok, O. Kurland, D. Carmel, F. Raiber, and G. Markovits. Predicting query performance by query-drift estimation. *ACM, TOIS*, 30(2):11, 2012.