# Assessing the Cognitive Complexity of Information Needs

Alistair Moffat
The University of Melbourne, Australia
ammoffat@unimelb.edu.au

Peter Bailey
Microsoft, Australia
pbailey@microsoft.com

Falk Scholer
RMIT University, Australia
falk.scholer@rmit.edu.au

Paul Thomas
CSIRO, Australia
paul.thomas@csiro.au

## ABSTRACT

Information retrieval systems can be evaluated in laboratory settings through the use of user studies, and through the use of test collections and effectiveness metrics. In a larger investigation we are exploring the extent to which individual user differences and behaviours can affect the scores generated by a retrieval system.

Our objective in the first phase of that project is to define information need statements corresponding to a range of TREC search tasks, and to categorise those statements in terms of task complexity. The goal is to reach a position from which we can determine whether user actions while searching are influenced by the way the information need is expressed, and by the fundamental nature of the information need. We describe the process used to create information need statements, and then report inter- and intra-assessor agreements across four annotators. We conclude that assessing the relative cognitive complexity of tasks is a complex activity, even for experienced annotators.

## 1. INTRODUCTION AND BACKGROUND

Information retrieval is big business, with billions of dollars of advertising revenue at stake in the web search context, and billions of dollars of software sales at stake in enterprise and desktop search markets. But comparing systems – to answer the question, "is System B (new) better than System A (old)?" – remains a challenging task, despite several decades of research activity. User studies have the advantage of reflecting the behaviours of actual users, and allow measurement of task-oriented performance criteria, such as completion rates and time taken, together with user-oriented criteria, such as satisfaction. But user studies have the disadvantage of expense, and are typically (but not always) carried out at relatively small scale, over dozens of subjects rather than tens of thousands. User studies are also problematic in terms of what might be called iterability – there is a sense of holding constant the majority of the possible experimental variables, so that only a few factors are measured during the experiment. If those fixed experimental variables are set incorrectly, it isn't easily possible to alter them and rerun the experiment. Large-scale search companies are the exception, and are able to carry out live testing by diverting a fraction of their query stream to "System B" while sending the majority to "System A".

The other major type of system evaluation is batch experimentation using a corpus, a set of provided topics, and a set of pre-judged relevance assessments. Systems are scored using numeric metrics that the researcher believes represent user needs or behaviours, to allow comparison using standard statistical techniques. This "Cranfield/TREC" evaluation style has the benefit of being repeatable – indeed, some would say too repeatable. But compiling the resources required is still costly, and most resource sets consist of hundreds of topics and partial (limited-depth) judgements, rather than thousands of topics and complete judgements. There are also issues to do with judgement quality, since two people, presented with the same information need and the same document, might (and often do) disagree over the responsiveness of that document to that topic.

Moreover, batch evaluations suffer from another significant risk: that by exploring system differences in the context of fixed topic sets, a much greater source of variability in behaviour is potentially being ignored. We are thus interested in the question: *to what extent is system effectiveness dependent on variations in user behaviour when using those systems?* That is, we wonder if the focus in batch evaluations on fixed queries and simple metrics as somehow being representative of "the (impartial) user" has meant that other issues have failed to be properly recognised.

We are not the first to investigate this phenomenon. The 1999 TREC "Query Track" examined sets of queries for topics, and the track coordinators Chris Buckley and Janet Walz concluded that [3]:

> *We've reaffirmed the tremendous variation that sometimes gets hidden underneath the averages of a typical IR experiment.*
>
> - *Topics are extremely variable*
> - *Queries dealing with the same topic are extremely variable. . . .*
> - *Systems were only somewhat variable.*

Our project encompasses exploration of such variability, and the desire to connect user experiences with the effectiveness quality determined by batch evaluations.

## 2. SOURCES OF VARIATION

Variation in evaluation comes from at least two sources besides search systems themselves: differences in metrics, and differences in user behaviour, either innate or due to task.


*ADCS'14,* November 27–28 2014, Melbourne, Victoria, Australia.

ACM 978-1-4503-3000-8/14/11 ...$15.00.
http://dx.doi.org/10.1145/2682862.2682874.

**Metrics** The choice of scoring regime is a key factor that affects batch evaluations. Traditional batch evaluations have employed metrics such as Reciprocal Rank (RR); Precision at depth $k$, for example P@10; and Average Precision, AP, computed as the per-query average of the precisions achieved at the location of each relevant document in the ranking. Scores for a metric are averaged over a set of topics to get an overall value, and sets of scores are compared using statistical techniques. But all of these metrics have failings of various sorts, and in response a range of further alternatives have been developed over the last decade, including Normalised Discounted Cumulative Gain, NDCG [6]; Rank-Biased Precision, RBP [8]; and Expected Reciprocal Rank, ERR [4].

Metrics are asserted to measure different aspects of performance, and to be sensitive in different ways. For example, AP and NDCG are regarded as being "deep" metrics, and are typically calculated to ranking depths 100 or even 1,000 (which is expensive in terms of judgement effort required); they are supposedly good for detecting fine differences in performance, as well as being predictive of other less-sensitive metrics. Precision at 10 is a "shallow" metric, and is both much cheaper to evaluate, and also better reflects the behaviour of a typical web search user. And so on. Moffat [7] provides further commentary on ways we can categorise effectiveness metrics.

Hence, a typical batch evaluation consists of building a "System B (new)" that implements some idea; running a set of queries against it and "System A (old)"; applying a suite of alternative metrics to the runs; computing statistical significance; and then (hopefully) adding some daggered "significant at $p < 0.05$" annotations to a table of mean effectiveness scores. That is, all users, and all possible queries that they might issue to address an underlying information need, are regarded as being "equal", and hence capable of being measured using the same technology. Then a range of metrics are used to carry out the measurement, so that all bets are covered.

There has however been a growing appreciation that users differ in the way in which they process query response pages. Variation in user behaviour will lead to variation in system effectiveness, and arguably our evaluations and metrics need to account for this.

**Goal and persistence** Moffat and Zobel [8] noted that users vary in their persistence: their desire to read summaries or documents, or to issue new queries. This is reflected in the parameter $p$ used in RBP. Moffat et al. [9] further noted that different searchers will have different goals, even for the same query or same information need, due to (for example) different background knowledge, or different learning styles. They introduce the notion of an "expected goal of search", their parameter $T$, and use it to shape predictions about what happens when that user is viewing a page; this more refined user model then leads to alternative effectiveness metrics. Moffat et al. carried out a user study that provided evidence supporting that hypothesis, bringing user and batch evaluations a step closer.

**Query text** The translation between what might be called "information seeking intent" and "query" is a source of variability when retrieval systems are evaluated. Users misspell words, use little or no punctuation in queries, and for the most part, don't type full sentences. So there is a potential mismatch between a query like "*What is the state song of Kansas?*" (topic 1840, 2002 QA Track), and the alternative query "*song kansas*", which may refer to a song by the rock band Kansas. The outcome of a batch comparison of two systems seems at least as likely to depend on which of these two queries is issued as it does on innate system performance [3].

**Task complexity** Clearly, different information-seeking tasks have different characteristics. Task complexity in particular is likely to influence behaviour, so should certainly influence our notion of effectiveness, and hence also influence our choice of metric.

Wu et al. [10] discuss the nature of information search, building on their earlier work investigating vertical search and user interaction [2], which employed three approaches to characterising the complexity of constructed tasks. They propose a cognitive complexity hierarchy for queries derived from a similar taxonomy of learning objectives presented by Anderson and Krathwohl [1]. They suggest that a spectrum of information needs should be considered. At the lowest level of their hierarchy, *Remember* queries represent the desire to recall a simple fact that is the response to a straightforward "when" "where" or "what"-type question, for example, "*what is the melting point of lead*", or "*where were the 1956 Olympics*". Other increasingly complex levels in their hierarchy include *Understand*, described by Wu et al. [10] as:

> *Constructing meaning from oral, written, and graphic messages through interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining.*

and *Analyse* [10]:

> *Breaking material into constituent parts, determining how the parts relate to one another and to an overall structure or purpose through differentiating, organizing, and attributing.*

Relating these definitions back to the $T$ parameter of Moffat et al. [9], it seems likely that different task complexities will result in users expecting to need different numbers of relevant documents before deciding that their information need has been satisfied.

**Planned research investigation** Our larger investigation aims to learn how much variation there is due to some of these factors, and what this means for Cranfield/TREC-style evaluations. Given tasks of different complexities, we plan to crowd-source a number of variant queries (*query text*), run these queries on a standard search engine, and investigate the resulting distribution of effectiveness scores. We also plan to ask for estimates of Moffat et al.'s $T$ (encapsulating both *goal* and *persistence*), to test whether this is a useful construct, and what it may mean for effectiveness measures.

# 3. PROCESS AND RESULTS

To investigate the effect of task complexity in a controlled manner, we need a set of labelled tasks. We now describe the first phase of our larger project, namely, categorising a set of TREC topic statements according to task complexity, and carrying out consistency checks to validate that work.

**Rewriting topics** A set of 180 TREC topics was extracted from three different tracks/years (see `http://trec.nist.gov/tracks.html`) to give a broad cross-section of information-seeking tasks. The three tracks selected were:

- Question Answering Track, 2002, 70 topics, 1824–1893;
- Robust Track, 2003, 60 topics selected from 303–610;[1]
- Terabyte Track, 2004, 50 topics, 701–750.

Each topic, including any TREC-provided narrative and description components, was then read by a *primary annotator*, who created the first draft of the *backstory*, a brief statement of information need designed to motivate the subsequent search, and intended to be used as part of a crowd-based data collection activity. The primary

[1] Half the topics selected for Robust 2003 were ones known to be difficult from previous years' TREC ad hoc tracks, hence the non-contiguous topic numbers.

Q02.1866, *Remember*; "What part of the eye continues to grow throughout a person's life?"

You went to an eye safety lecture as part of starting a new job in a factory. The presenter said that there is one part of the eye that continues to grow throughout a person's life. What is that part?

R03.397, *Understand*; "Automobile recalls"

Your car recently got recalled by the manufacturer so that a small defect could be repaired. You start wondering how often cars get recalled, and what type of major or minor reasons are behind such recalls.

T04.706, *Analyse*; "Controlling Type II Diabetes"

Your doctor recently told you that you are at risk of type II diabetes. You decide to do your own research, to find out the relationship between weight, exercise, and diabetes; other ways of keeping triglycerides, cholesterol and blood pressure in normal ranges; and controls such as determining blood sugar levels.

Figure 1: Three TREC topics from different tasks in different years, together with the primary assessment of the topic type, and with the backstory generated for each one. A total of 180 information need statements were generated from these three TREC Tracks.

*Number*: 706

Controlling type II diabetes

*Description*:
What are methods used to control type II diabetes?

*Narrative*:
Items containing such controls as determining blood sugar levels and keeping triglycerides, cholesterol and blood pressure in normal ranges are relevant. Mention of mild to moderate weight loss, regular exercise and learning new behaviors and attitudes, medications is relevant.

Figure 2: Topic 706 from the TREC 2004 Terabyte Track.

annotator was able to explore background information using on-line sources as part of that process.

Figure 1 shows three examples of topics from these tracks, and the backstory that was constructed for each one. The Robust and Terabyte Track topic descriptions provided by TREC contain considerable detail about the nature of documents that can be categorised as relevant, and we sought to capture those requirements, or the bulk of those requirements, in the corresponding backstory. Figure 2 shows the original TREC presentation of one of the topics shown in Figure 1. Each backstory was personalised, via the use of "you" and hypothetical friends and family members; and many were written as "you decide" or "you wonder" passive suggestions, rather than as imperative "find" or "what" questions. We want our eventual subjects to empathise with the information need, and treat it as a personal search rather than an impersonal one.

The QA Track topics were in some ways harder to deal with than the Terabyte and Robust ones. They are presented by TREC as unadorned questions, for example, "*What part of the eye continues to grow throughout a person's life?*" (topic 1866). But we were conscious that simply posing those questions in a crowd-sourcing interface might encourage the subjects to regard the question as being the query they were expected to issue, rather than an expression of an information need. So the QA topics are also presented with a backstory, and when possible, we used pronouns or other indirect references to the query subject, in order to remove the temptation for crowd-sourced subjects to cut and paste the final "*what?*" question as their query. We also wish to present the various backstories without including obvious cues such as length or voice, which might affect the way that the experimental subjects react to them, a further reason to extend the QA ones.

**Task complexity categorisation** Once the backstory was created, the primary annotator also assigned one of the task complexity types described by Arguello et al. [2], and motivated in Section 2, that is, *Remember*, *Understand*, and *Analyse*. It was relatively straightforward to identify the information seeking tasks that fitted the *Remember* category, and considerably harder to differentiate between the other two types. Overall, where topics required the production of list of things, even if relatively complex and sourced from different pages, we ended up tending to make them *Understand*; where topics required synthesis of disparate information, and eventual summary, or balancing of competing viewpoints and opinions, we categorised them as *Analyse*.

After this initial labelling round, we had 71 *Remember* tasks; 79 *Understand* tasks; and 30 *Analyse* tasks. Each annotator then provided a category annotation on all 180 topics (including their own original assigned topics) when presented in a randomised order, viewing only the topic's backstory, and using that as the basis for a fresh decision as to the task category. In this labelling round, each of the four annotators provided a label (*Remember*, *Understand*, or *Analyse*), optionally qualified with a "?" to indicate uncertainty.

The full set of 180 query topic backstories and corresponding task complexity labels is available on request.

**Measuring agreement** To simplify agreement statistics, we collapsed the second-round annotations into the three primary categories, dropping any "?" indicators. We assessed inter-annotator agreement using Fleiss' kappa [5], a test statistic that accommodates multiple raters and corrects for agreement by chance. This statistic is calculated over the entire 180 topic categories, each labeled by all 4 annotators, ignoring the originally assigned label, and using only the secondary opinion labels. Intra-annotator agreement is measured using Cohen's kappa [5], between the annotator's original label and their secondary opinion on just those 45 topics for which they created the backstory.

**Results** The inter-annotator agreement values are reported in Table 1, including overall agreement across all three categories, and also broken down by the individual categories. Agreement statistics are notoriously difficult to interpret, so we remain reluctant to conclude that the particular agreement levels "mean" something. The

| Category | Fleiss' kappa | $z$ | $p$-value |
|---|---|---|---|
| *All categories* | 0.664 | 30.0 | $< 0.0001$ |
| *Remember* | 0.907 | 29.8 | $< 0.0001$ |
| *Understand* | 0.563 | 18.5 | $< 0.0001$ |
| *Analyse* | 0.456 | 15.0 | $< 0.0001$ |

Table 1: Inter-annotator greement measured by Fleiss' kappa.

| Annotator | Cohen's kappa | $z$ | $p$-value |
|---|---|---|---|
| A | 0.608 | 6.18 | $< 0.0001$ |
| B | 0.645 | 5.85 | $< 0.0001$ |
| C | 0.794 | 7.63 | $< 0.0001$ |
| D | 0.920 | 7.39 | $< 0.0001$ |

Table 2: Intra-annotator agreement measured by Cohen's kappa.

intuition for Fleiss' kappa is that it assesses the degree to which the annotators' agreement differs from what would have happened if they had provided labels completely at random. We observe that the *Remember* category achieves the highest degree of agreement across the four annotators, followed by the *Understand* category, and that the *Analyse* category is least likely to be agreed upon. The $z$ scores and $p$-values show the outcomes of statistical significance tests of the null hypothesis that the population level of kappa is zero (that is, the level of agreement is the same as expected by chance).

Intra-annotator agreement values are reported in Table 2. Annotators are listed anonymously as A, B, C, and D, in increasing order of self-agreement. We note there is a spread of self-agreement in the Cohen's kappa scores, which may be partially explained by differences in elapsed time between the original labelling and re-labelling by the same judge. Other factors may also have contributed to the spread, including time taken to complete the task.

Using a simple majority label assignment rule, of the 180 topics, 24 had no majority label (that is, did not achieve three or more annotators agreeing on the label). Of these 24 topics, seven had at least one *Remember* label, while the remaining 17 were equally split 2:2 between *Understand* and *Analyse*. We carried out a re-labelling exercise over the seven topics which had at least one *Remember* topic, to see if we could produce majority labels. This exercise was successful. One especially interesting topic ("*What is the major crop grown in Arizona*") provoked three of the four annotators to adjust their label, and the author of the original topic backstory ultimately moved in to a minority, assigning it to *Understand*, while the remaining three annotators agreed on *Remember*. The cause of the confusion was the backstory text, which originally read: "*You live in Texas, and are familiar with the sight of cotton crops. While planning a road trip to the Grand Canyon you start wondering what other agricultural commodities you will see along the way, including what the major crop grown in Arizona is*". The confusion over labelling was caused by the ambiguity over how many states were being traversed, and the use of the word "including". That backstory will be changed to "*You live in New Mexico, and are familiar with the sight of sorghum crops. While planning a road trip to the Grand Canyon you start wondering what will be the main crop you see in Arizona*".

Examining the remaining split-labeled topics, two of the annotators were approximately twice as likely to assign *Analyse* labels to these topics than the other two annotators. It was also the case that 6 of the 17 topics had at least one "?" qualifier assigned, indicating a degree of uncertainty in the label assignment. This represented a qualifier rate twice the average for the topic set as a whole. (Two of the annotators, A and C did not add any qualifiers; annotator B assigned 10 qualifiers; and annotator D assigned 20.) We hoped to use the original annotator's label as a tie-breaker for these 17 topics. But further analysis revealed that in around half such cases the original annotator changed their label, providing continued evidence of ambiguity in the interpretation of *Understand* and *Analyse*.

# 4. CONCLUSIONS

Based on the agreement analysis described above, we conclude that *Remember* tasks are relatively straightforward to characterise by cognitive complexity *a priori*, as indicated by the high level of agreement among four annotators. We are less confident that such a clear division is possible between the *Understand* and *Analyse* complexity levels. It may be that as annotators we have different concepts of how complex a particular information need will be, or that more detailed guidelines need to be written describing boundary conditions between levels, or that predicted cognitive complexity is indeed a continuum rather than separable into levels. Should we need to generate equivalent complexity labels in the future, we might use the instances of topics with non-majority labels as sources for guideline development to improve label consistency.

In the main part of this project we plan to provide the topic backstories to crowd-sourced workers, and ask them to tell us how many queries and documents they would expect to require to satisfy the information need, as well as their starting queries. We will use our majority labels as to the corresponding cognitive complexity for the tasks as a dimension in analysis of the collected data. Our goal is to understand whether different levels of cognitive complexity in tasks lead to systematic differences in the expectations of searchers about the number of queries and number of documents needed, and to the first queries they then ask; or whether any such differences are primarily due to individual variation in the searchers themselves.

**Acknowledgment** This work was supported by the Australian Research Council's *Discovery Projects* Scheme (projects DP110101934 and DP140102655).

# References

[1] L. W. Anderson and D. A. Krathwohl. *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York, 2001.

[2] J. Arguello, W.-C. Wu, D. Kelly, and A. Edwards. Task complexity, vertical display and user interaction in aggregated search. In *Proc. SIGIR*, pages 435–444, 2012.

[3] C. Buckley and J. Walz. The TREC-8 query track. In *Proc. TREC*, 1999. NIST Special Publication 500-246.

[4] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proc. CIKM*, pages 621–630, 2009.

[5] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378, 1971.

[6] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Information Systems*, 20(4):422–446, 2002.

[7] A. Moffat. Seven numeric properties of effectiveness metrics. In *Proc. AIRS*, pages 1–12, 2013.

[8] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Information Systems*, 27(1):2:1–2:27, 2008.

[9] A. Moffat, P. Thomas, and F. Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proc. CIKM*, pages 659–668, 2013.

[10] W.-C. Wu, D. Kelly, A. Edwards, and J. Arguello. Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. In *Proc. IIiX*, pages 254–257, 2012.