On obtaining effort based judgments for Information retrieval

Manisha Verma, Emine Yilmaz, Nick Craswell
UCL UCL Microsoft

Motivation

1. **Log based evaluation**
   - Relevance determined implicitly or by clicks
   - Large scale collection of user data
   - Suffers from several (click, presentation etc) biases.

2. **Batch evaluation**
   - Relevance labels assigned by trained judges
   - Smaller test collections
   - Simple assumptions about real information needs

These two forms of evaluation **often do not completely agree** with each other ([1] and [2])
They **agree** with each other only **when there is a significant gap** in quality of the systems compared ([3] and [4]).

Trained Judges

End Users

- At present, ***relevance is primary factor*** for judging documents. It does not consider **'User effort'** (Yilmaz et al. [5])

- A **judge can spend a lot of time** evaluating correctness of document for a given query. An **impatient user may not spend as much time** studying the document!.

User Model

1. When users first access the page, they ***quickly scan it*** to determine portions relevant to the query.

   FINDABILITY

2. This is ***followed by reading*** these paragraphs/snippets.

   READABILITY

3. Finally, user ***focusses on understanding*** these nuggets of information.

   UNDERSTANDABILITY

Contributions

1. Identify ***factors that characterize*** user effort.

2. Conduct experiments to ***obtain explicit judgments*** for these factors.

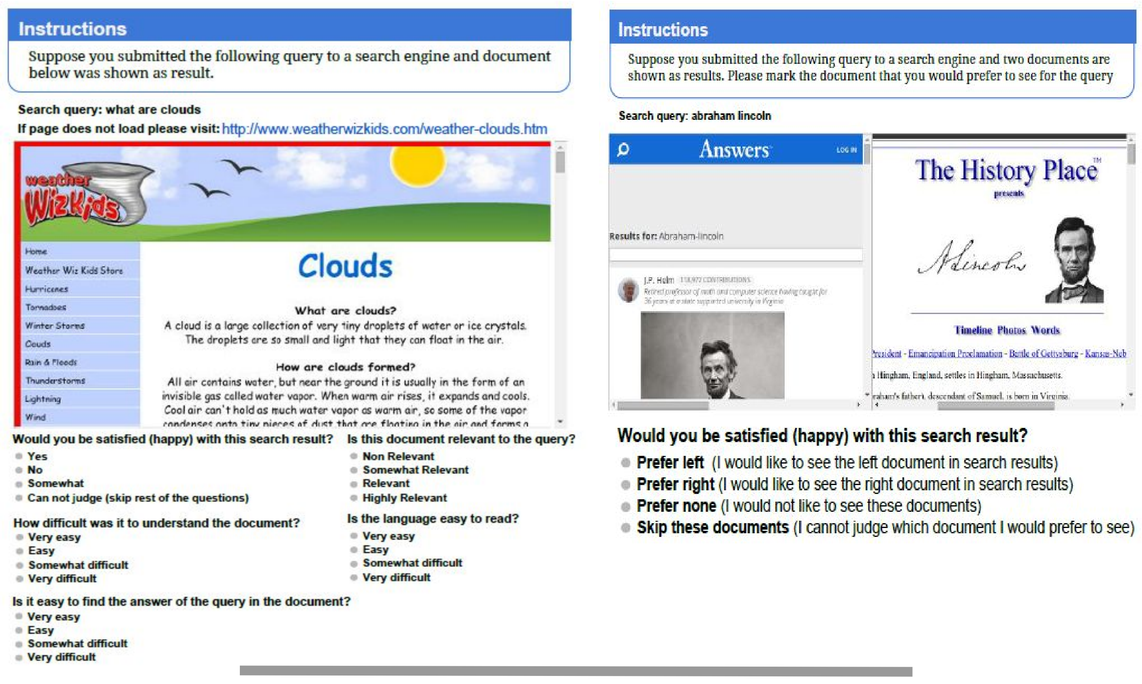
3. Finally analyse the ***effect of incorporating effort*** into retrieval evaluation.

Methodology

- Collect **effort based (explicit) judgments** for each document for above parameters.

- Study user preferences
  - **Control for relevance**: Collect user preferences with side-by-side comparison for documents of ***same relevance grade***.

Judging Interfaces



Results

Factors important for User Satisfaction

FACTOR	p-Value
Findability^{+}	0.003
Readability^{-}	0.364
Understandability^{+}	0.054
Relevance^{+}	0

Effort and Preference Agreement

FACTOR	p-Value
Findability	0.60*
Readability	0.51
Understandability	0.51
Relevance	0.72*

Features

Text Features	
avgSumChar	Avg #chars in summary
docCLI	CLI Index of document
docWords	#words in document
qTermsInTitle	#query terms in title
sumWords	#words in summary
tRatio	Fraction of #words and #tags in html

Structure Oriented Features	
fTable	Fraction of Tables
maxWinPos	Max window pos with all query terms
qWinO	Fraction of outlinks with query terms
fBoldItalics	Fraction of bold, italics and strong
fImg	Fraction of images
minWinPos	Min window pos with all query terms
countH	#Headings with query terms

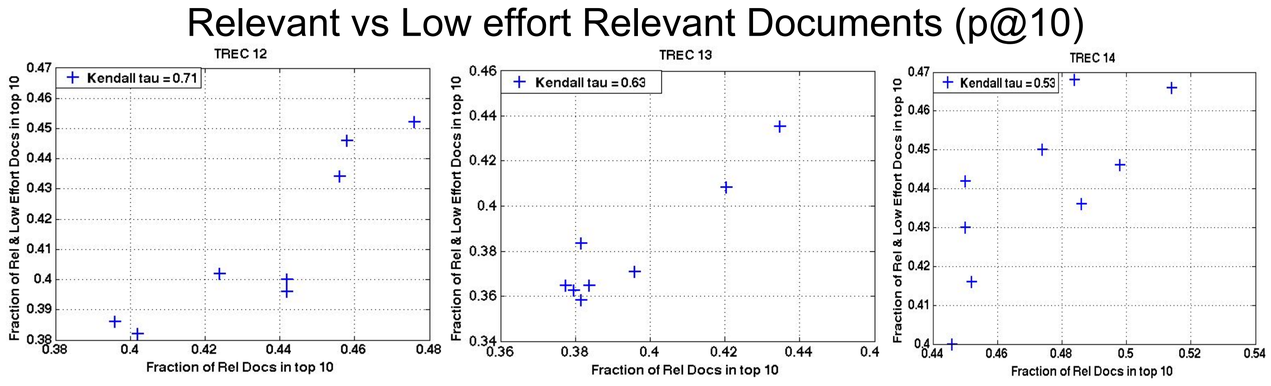
Findability Prediction

FEATURE	p-val	FEATURE	p-val
fTable^{-}	0.00	minWinPos^{+}	0.00
avgSumChar^{-}	0.01	meanPosOut^{+}	0.01
docCLI^{-}	0.02	sumWords^{+}	0.02
maxWinPos^{-}	0.04	fImg^{+}	0.02

Relevance Prediction

FEATURE	p-val	FEATURE	p-val
tRatio^{-}	0.01	qTermsInTitle^{+}	0.01
countH^{-}	0.02	qWinO^{+}	0.01
maxWinPos^{-}	0.04	fBoldItalics^{+}	0.02
docWords^{-}	0.04	fImg^{+}	0.04

Relevant vs Low effort Relevant Documents (p@10)



References

1. W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? In Proc. SIGIR, Athens, Greece, 2000.
2. A. H. Turpin and W. Hersh. Why batch and user evaluations do not give the same results? In Proc. SIGIR, New Orleans, USA, 2001.
3. J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In Proc. SIGIR, Salvador, Brazil, 2005.
4. A. Al-Maskari, M. Sanderson, P. Clough, and E. Airio. *The good and the bad system: Does the test collection predict users' effectiveness?* In Proc. SIGIR, 2008.
5. E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. Bailey. *Relevance and effort: An analysis of document utility*. In Proc. CIKM, 2014.