User Models, Metrics and Measures of Search

A Tutorial on the C/W/L Evaluation Framework

Leif Azzopardi
University of Strathclyde
Glasgow, UK
leifos@acm.org

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

Paul Thomas
Microsoft
Canberra, Australia
pathom@microsoft.com

Guido Zuccon
The University of Queensland
Brisbane, Australia
g.zuccon@uq.edu.au

ABSTRACT

Evaluation is central to Information Retrieval, and is how we compare the quality of systems. One important principle of evaluation is that the measured score should reflect the user's experience with the system. Hence, there should be direct connection between how user's interact with the system and the characteristics of the metric. In this tutorial we introduce the C/W/L approach to user modeling and show how different user models lead to different metrics. We then describe the recent innovations and approaches to evaluation that it has facilitated. The tutorial is presented as a mix of on-line synchronous lecture, pre-recorded in-depth videos, and hands-on activities using the C/W/L toolkit for particpant's own evaluation tasks. A followup consultation session is also provided, to allow extended questions and individual discussion with the four presenters.

ACM Reference Format:
Leif Azzopardi, Alistair Moffat, Paul Thomas, and Guido Zuccon. 2021. User Models, Metrics and Measures of Search: A Tutorial on the C/W/L Evaluation Framework. In *Proceedings of the 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '21), March 14–19, 2021, Canberra, ACT, Australia.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3406522.3446049

1 MOTIVATION

Effectiveness evaluation has played a central role in the development of information retrieval systems [10]. Many effectiveness metrics have been proposed over the years, with the more recent ones employing multi-valued and/or discounted relevance values (Discounted Cumulative Gain, DCG [4], and Rank Biased Precision, RBP [6]); and/or the cost (or time) associated with viewing result items (Time Biased Gain, TBG [11]); and/or the way in which users adapt their interactions according to their goals and constraints (INST [8], the Bejewelled Player Model, BPM [13], and Information Foraging Theory IFT [1]).

With such a diversity of metrics to choose from, there are many questions to be considered: what assumptions do they make, what do they measure, are they measuring the same or different quantities, what units are the measurements in, which one should we be using, and what tools/scripts are available to run them?

Underlying most metrics, sometimes explicit and sometimes implicit, is a User Browsing Model (UBM) that encodes how people interact with the results presented; with different metrics making different assumptions about how people browse the search results and derive utility from them. Under the C/W/L framework [9] it is possible to describe a range of both traditional and recently proposed models and metrics as being variations of a single overall approach. This has a number of advantages:

- measurements taken are in the same units, and thus can be compared between metrics, meaning that (for example) the estimated Expected Utility (expected rate of gain) as given by RBP is directly comparable to the estimated Expected Utility as given by TBG, BPM, INST, IFT, and so on;
- the C/W/L framework enables the estimation of a series of measurements beyond Expected Utility, including Expected Total Utility, Expected Total Cost, and Expected Search Depth;
- some measurements can be compared directly to observables, for example, Time Spent, Last Item Examined, and so on, and hence allow parameter estimation to be undertaken; and
- new metrics can be encoded by instantiating an appropriate User Browsing Model formally defined by a Conditional Continuation Probability function.

Taken together, these advantages mean the C/W/L framework provides an extensible and versatile basis for measuring different aspects of retrieval performance under different user modeling assumptions, with the flexibility provided by the fourth point perhaps of greatest importance.

2 LEARNING OBJECTIVES

As a result of attending this tutorial, participants will be able to:

- formally define the C/W/L framework and how to calculate the Expected Utility (also known as expected rate of gain) of a result list;

CHIIR '21, March 14–19, 2021, Canberra, ACT, Australia

ACM ISBN 978-1-4503-8055-3/21/03…$15.00
https://doi.org/10.1145/3406522.3446049

- define the User Browsing Models (continuation functions) of different metrics;
- show how new metrics can be defined via the UBM;
- explain how the C/W/L framework can be extended to produce different measurements;
- describe further extensions to the C/W/L framework to include snippet-based and session-based measurements;
- show how the C/W/L framework can be used to evaluate results and analyze Continuation functions, Weighting functions, and Last Likelihood Functions;
- use the "`cwl_eval`" toolkit [2] to perform TREC-like evaluations on typical IR system experimental outputs.

3 TUTORIAL FORMAT

Structure The tutorial consists of:

- an on-line live presentation of approximately 120 minutes, to describe the overall structure and principles of the C/W/L framework, and to illustrate a range of traditional and more recent metrics that fit it;
- access to a range of supporting pre-recorded videos that "zoom in" on particular topics, to allow attendees to review more complex material at their own speed and in their own time; and
- a consultation/workshop session of approximately 120 minutes, in which attendees are able to ask questions about the lecture and video material, and are also able to discuss their own particular evaluation tasks and requirements with the presenters.

Content The on-line session focuses on a brief introduction to evaluation in Information Retrieval to provide the necessary context, before defining the C/W/L framework [7, 9]. During this session it is explained how standard and commonly used IR retrieval measures can be defined within the framework, such as Precision, Average Precision, Discounted Cumulative Gain (DCG) [4] and Rank Biased Precision (RBP) [6]); and how the framework extends beyond relevance to focus on gain (utility).

Extensions to the C/W/L framework will also be described to explain how it can be used to compute more than the Expected Utility (the rate at which gain is acquired), such as *Expected Total Utility*, the gain accumulated from the whole list; *Expected Cost* per item inspected; *Expected Total Cost*, the cost incurred in examining the results list; and *Expected Depth*, the number of items a searcher examines given the user model encoded within the metric.

We then consider newer metric proposals, such as: Time Biased Gain (TBG) [11], where the UBM is dependent on time; INST [8] and INST-BA [5], where the UBM is dependent on gain; the Bejewelled Player Model (BPM) [13] and Information Foraging Theory (IFT) [1]), where the UBM is dependent on the costs, gains, and different constraints; and Data Driven Metrics (DDM) [3], where the UBM is derived directly from data.

We conclude by discussing how these metrics have been evaluated and validated against satisfaction, behaviors and observed data [3, 9, 12, 14, 15], considering further extensions to the C/W/L framework, including adding in snippets, and evaluation across sessions. Participants will be also be shown how to perform TREC-like evaluations using the "`cwl_eval`" tool [2], and how to implement their own metrics.

Audience The intended audience is predominately graduate students in Information Retrieval interested in modeling user behavior and measuring search performance – in other words, wishing to better understand IR metrics. The tutorial is also relevant to practitioners wanting to know how to model and measure how people interact with their systems and applications using a formal framework. This course would be particularly valuable for participants looking to: (1) provide a theoretical underpinning to the design, development and evaluation of their algorithms, interfaces, and applications; (2) reason about how the system influences behaviors and performance; and (3) ground and inform experimentation through measurement.

REFERENCES

[1] Leif Azzopardi, Paul Thomas, and Nick Craswell. Measuring the utility of search engine result pages: An information foraging based measure. In *Proc. SIGIR*, pages 605–614, 2018.

[2] Leif Azzopardi, Paul Thomas, and Alistair Moffat. cwl_eval: An evaluation tool for information retrieval. In *Proc. SIGIR*, SIGIR '19, 2019.

[3] Leif Azzopardi, Ryen W. White, Paul Thomas, and Nick Craswell. Data-driven evaluation metrics for heterogeneous search engine result pages. In *Proc. CHIIR*, pages 213–222, 2020.

[4] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

[5] Alistair Moffat and Alfan Farizki Wicaksono. Users, adaptivity, and bad abandonment. In *Proc. SIGIR*, 2018.

[6] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27, 2008.

[7] Alistair Moffat, Paul Thomas, and Falk Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proc. CIKM*, pages 659–668, 2013.

[8] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. INST: An adaptive metric for information retrieval evaluation. In *Proc. Aust. Doc. Comp. Symp.*, pages 5:1–5:4, 2015.

[9] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Trans. Inf. Syst.*, 35(3):24:1–24:38, 2017.

[10] Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

[11] Mark D. Smucker and Charles L.A. Clarke. Time-based calibration of effectiveness measures. In *Proc. SIGIR*, pages 95–104, 2012.

[12] Alfan Farizki Wicaksono and Alistair Moffat. Metrics, user models, and satisfaction. In *Proc. WSDM*, pages 654–662, 2020.

[13] Fan Zhang, Yiqun Liu, Xin Li, Min Zhang, Yinghui Xu, and Shaoping Ma. Evaluating web search with a bejeweled player model. In *Proc. SIGIR*, pages 425–434, 2017.

[14] Fan Zhang, Jiaxin Mao, Yiqun Liu, Xiaohui Xie, Weizhi Ma, Min Zhang, and Shaoping Ma. Models versus satisfaction: Towards a better understanding of evaluation metrics. In *Proc. SIGIR*, pages 379–388, 2020.

[15] Yuye Zhang, Laurence A. F. Park, and Alistair Moffat. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval*, 13(1):46–69, February 2010.