comprehensive answer. This may be of particularly great help for those searchers having poor information or media literacy. This is of obvious importance in many situations: e.g., education, medical information, and search for topics "that matter". Some special domains, such as patent search and evidence based practices in medicine, have clearly prescribed a particular information seeking process in great detail. Here building a systems to support (and enforce) this process is of obvious value.

4.9 Interaction, Measures and Models

Gianmaria Silvello (University of Padova, IT), Leif Azzopardi (University of Glasgow, GB), Charlie Clarke (University of Waterloo, CA), Matthias Hagen (Bauhaus–Universität Weimar, DE), Robert Villa (University of Sheffield, GB)

A common framework for user interaction models and a common framework in which to place evaluation measures (i.e., the units of measurement) should be consistent but does not yet exist. Current measures are not comparable as the units used are not clearly defined in terms of real-world outcomes, and vary between measures. Since most measures encode some form of user behaviour as an underlying user interaction model, having measures that use the same unit of measure would enable comparisons between different user interaction models across different systems. As well as making it possible to compare between measures themselves (opposed to viewing them independently in different units).

4.9.1 Motivation

The main goal is to enable assessment of the performances of the system as a whole or specific components in particular. For that we need a repeatable way to say that a system is better than another on a gain base (utility, usefulness, happiness, ...). Ideally, the effect of user attributes that are not salient to the evaluation itself should be minimized (e.g. "what the user had for breakfast"). The measures should be comparable; that is, defined using the same units (i.e. gain, cost, or gain/cost). We would also like to be able to determine the effects of the interface and interaction on the actual performance.

4.9.2 Proposed solution

Integrate the interaction with an IR interface into the measures, e.g. in a TREC- style evaluation, individual IR systems may submit conventional ranked lists. Systems can then be evaluated based on different models of user interfaces or interactions. To extend TREC-style evaluations to accommodate more realistic interfaces, individual systems might submit responses to a variety of user actions, which would then be evaluated across more complex and detailed interfaces and interaction models.

One possible solution would be to decompose measures into components: Interaction model (I) (traditionally: when the user stops) Gain model (G) (traditionally: number of viewed relevant docs) Cost model (C) (traditionally: number of viewed docs with unit costs) An evaluation measure could then be parameterized by the components as M(I,G,C).

An interaction model might be characterized by a sequence of states and for each state some specific interaction with the system taken; potentially depending on the intent and task

of the user (e.g., a recall oriented task). The interface of the system could be encoded in the cost model. The gain would model the documents returned to the user (e.g., degree of relevance)

For example, we can deconstruct DCG into the three main components outlined above: the interaction model is "defined" by the discounting function, the gain model is how we sum up the weights of viewed (relevant) documents and the cost model is the number of viewed documents (with a fixed cost for each document). This means that we can fix the gain and the cost models while changing the interaction model still being able to compare measurements.

We could define an idealized interaction between the user and the system (including its interface). Idealized in this case would mean the optimal behavior where users are able to make decisions towards the best possible gains. System comparison based on such idealized interactions seem to be much more reasonable and comparable than based on arbitrary and possibly sub optimal decisions. This approach would also enable us to drop from the models a number of parameters that are difficult to estimate, such as click and query reformulation probabilities.

References

1 Allan, J., Croft, B., Moffat, A., and Sanderson, M. (2012). Frontiers, challenges, and opportunities for information retrieval: *Report from SWIRL 2012 the second strategic workshop on information retrieval in Lorne. SIGIR Forum 46:1* (May 2012), pp. 2–32, ACM Press, New York, NY, USA . http://doi.acm.org/10.1145/2215676.2215678

2 Azzopardi, L. (2009). Usage based effectiveness measures: monitoring application performance in information retrieval, *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*, pp. 631–640, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=1646034

3 Azzopardi, L. (2011). The economics in interactive information retrieval. *Proceedings of the 34th international ACM SIGIR International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'11)*, pp. 15–24, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=2009916.2009923

4 Azzopardi, L., Kelly, D., and Brennan, K. (2013). How query cost affects search behavior. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*, pp. 23–32, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=2484049

5 Azzopardi, L., Järvelin, K., Kamps, J. and Smucker, M. D. (2011). Report on the SIGIR 2010 workshop on the simulation of interaction, *SIGIR Forum 44:2 (December 2010)*, pp. 35-47, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=1924475.1924484

6 Baskaya, F., Keskustalo, H., and Järvelin, K. (2012). Time drives interaction: Simulating sessions in diverse searching environments. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*, pp. 105–114, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=2348301

7 Belkin, N., Dumais, S., Kando, N. and Sanderson, M. (2012). N*II Shonan meeting on Whole-Session Evaluation of Interactive Information Retrieval Systems, Held October 2012*. To appear. http://www.nii.ac.jp/shonan/seminar020/

8 Carterette, B. (2011). System effectiveness, user models, and user utility: a conceptual framework for investigation, *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'11)*, pp. 903–912, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=2009916.2010037

9 Carterette, B., Kanoulas, E. and Yilmaz, E. (2012). Incorporating variability in user behavior into systems based evaluation. *Proceedings of the 21st ACM international Conference*

on Information and Knowledge Management (CIKM'12), pp. 135–144, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=2396761.2396782

10 Sakai, T. and Dou, Z. (2013). Summaries, ranked retrieval and sessions: a unified framework for information access evaluation, *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*, pp. 473–482, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=2484028.2484031

11 Smucker, M. and Clarke, C. (2012). Time-based calibration of effectiveness measures, *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*, pp. 95–104, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=2348283.2348300

12 Smucker, M. and Clarke, C. (2012). Modeling User Variance in Time-Biased Gain, *Proceedings of the Sixth Symposium on Human-Computer Interaction and Information Retrieval (HCIR'12)*, Article 3, 10 pages, ACM Press, New York, NY, USA. http://dl.acm.org/citation.cfm?id=2391224.2391227