Building and Using Models of Information Seeking, Search and Retrieval

Full Day Tutorial

Leif Azzopardi
University of Glasgow
Glasgow, United Kingdom
Leif.Azzopardi@glasgow.ac.uk

Guido Zuccon
Queensland University of Technology
Queensland, Australia
g.zuccon@qut.edu.au

ABSTRACT

Understanding how people interact with information systems when searching is central to the study of Interactive Information Retrieval (IIR). While much of the prior work in this area has either been conceptual, observational or empirical, recently there has been renewed interest in developing mathematical models of information seeking and search. This is because such models can provide a concise and compact representation of search behaviours and naturally generate testable hypotheses about search behaviour.

This full day tutorial focuses on explaining and building formal models of Information Seeking and Retrieval. The tutorial is structured into four sessions. In the first session we will discuss the rationale of modelling and examine a number of early formal models of search (including early cost models and the Probability Ranking Principle). Then we will examine more contemporary formal models (including Information Foraging Theory, the Interactive Probability Ranking Principle, and Search Economic Theory). The focus will be on the insights and intuitions that we can glean from the math behind these models. The latter sessions will be dedicated to building models that optimise particular objectives which drive how users make decisions, along with a how-to guide on model building, where we will describe different techniques (including analytical, graphical and computational) that can be used to generate hypotheses from such models. In the final session, participants will be challenged to develop a simple model of interaction applying the techniques learnt during the day, before concluding with an overview of challenges and future directions.

This tutorial is aimed at participants wanting to know more about the various formal models of information seeking, search and retrieval, that have been proposed in the literature. The tutorial will be presented at an introductory level, and is designed to support participants who want to be able to understand and apply such models, as well as to build their own models.

SIGIR'15, August 09-13, 2015, Santiago, Chile.
ACM 978-1-4503-3621-5/15/08.
DOI: http://dx.doi.org/10.1145/2766462.2767874.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval:Search Process; H.5.m [**Information Interfaces and Presentation**]:Miscellaneous

General Terms

Theory, Experimentation, Economics, Human Factors

Keywords

Search Behaviour; User Models; Retrieval Strategies; Evaluation

1. INTRODUCTION

The tutorial will be structured in four main parts: (1) Why Model and Early Models (1960-1980) (2) Contemporary Models (1980+), (3) Model building, and, (4) Making a model (as an hands-on activity for the participants). The day will conclude with a discussion of challenges and future directions. It should be noted that in the field of Information Retrieval we have developed lots of models - mainly retrieval models. However, this tutorial will focus on building models for interactive information seeking and retrieval, and so focus on modelling the user's interactions with the interface/system in order to characterise search behaviours. Such models are rarer, as they tend to be more complicated, and require a greater appreciation of how people interact with systems. Indeed, developing such models has been hailed as a (somewhat) grand challenge for Interactive Information Retrieval [15]. Furthermore, such models are becoming increasingly important as they form the basis of many performance measures, provide a better understanding of how people interact with such systems, and a way to evaluate the utility of various features/interactions.

1.1 Why Model and Early Models

In this part of the tutorial, we will develop the motivation and rationale for developing formal (by which we mean mathematical) models of Information Seeking and Retrieval (ISR). To begin, we will discuss and describe various conceptual and descriptive models of ISR, including Bates' Berry Picking Model [13] and the ISR framework proposed by Ingwersen and Järvelin [27], along with other such models (e.g. [14, 16, 17, 29, 28, 30, 23, 46]). This will provide the background for the tutorial where we point out the limitations of such models and motivate the need to develop models that are not only descriptive in nature, but predictive and crucially explanatory. To set the context, we will discuss the qualities of information, the poverty of attention, information as a good, and how information can be valued. This is important because information seeking is intrinsically embedded within a task. Once motivations and context are provided, we will then examine a number of early formal models (1960-1980s) that have been developed using Decision Theory, Economics and Cost-Benefit frameworks [1, 36, 22, 35]; the most notable being the Probability Ranking Principle (PRP) [35]. The following topics will be covered:

1. Types of Models (Conceptual, Descriptive, Formal)
2. An overview of Conceptual models
3. What is a model? Why formal, why mathematical?
4. Why formally model ISR? Why we need theory? At least, in theory!
5. The benefits and costs of formally modelling ISR
6. Why don't we have lots of formal models for ISR?
7. A discussion on the lack, and lack of usage, of formal models for ISR
8. The Information Age, Poverty of Attention, Mathlus' Law [21, 33]
9. Information as Good and the Value of Information [44]
10. Decision Making and Optimization Models [26, 31]
11. Optimality and Rationality [40, 39, 42]
12. Cooper's Model of Searching [22]
13. Probability Ranking Principle [35]

1.2 Contemporary Models

The second part of the tutorial will focus on three contemporary formal models. First we shall show how the PRP can be extended to consider interaction by presenting the Interactive Probability Ranking Principle [25]. Next, we shall delve into Information Foraging Theory (IFT), which has become very popular and is often used to motivate experiments, but yet infrequently used in terms of modelling and prediction. Here, we will focus on the patch model, and explain the different implications of the model/theory, and how user behaviour is expected to change under varying circumstances. We will explain how these implications can be generated through a graphical analysis (and for the more math savvy, how they can be analytically derived). Then, we will turn our attention to the initial model of Search Economic Theory (SET) [3] and explain how the model was developed, and how it is possible to graphically and computationally explore how these models can be used to make hypotheses about search behaviour. We will then show how the model was refined to include more variables (and thus become more realistic) [7].

A summary of the topics we shall cover in this section are:

1. Interactive Probability Ranking Principle [25]
2. Sensemaking and Cost Structures [37]
3. Optimal Foraging Theory [41]
4. Information Foraging Theory [32, 20, 34, 37, 38]:
   - Information Scent Model
   - Information Diet Model
   - Information Patch Model
   - Charnov's Marginal Value Theorem [19]
5. Economic Theories:
   - Optimal Search Behavior and the Pandora's Box Problem [42, 45, 44]
   - Optimal Amount of Information (when to stop and deciding at the margin) [18]
   - Search Economic Theory [3, 5, 6, 7]

During this session activity sheets will be given to participants to draw hypotheses from various models.

1.3 A Guide to Modelling

Part three will focus on optimisation models, in general, how to build such models, and how to use them to generate hypotheses using various methods (analytical, graphical and computational) [43, 31]. The essence of such models is to develop a cost function and a gain function given the different interactions that users can perform. Given these functions, it is then possible to establish under what conditions cost is minimized and/or gain is maximized. This section will explore the following topics:

- What choices are available? What are the limitations?
- What are the interactions, costs and benefits?
- What will vary? Parameters?
- Define the problem and the goal
- Construct a cost and gain function
- Solve, plot, compute
- Draw inferences

To help illustrate how to undertake the model building process, we will use an example based on the simple scenario of finding the first highly relevant document in a result list. We will describe how we can characterize the problem, enumerate the different variables, and show how the different variables impact the overall cost and the design of the interface. If time permits, then we will also include some notes on developing simulations to explore such models (i.e., computational approaches).

1.4 Practical Session

The final part of the tutorial will be dedicated to building a model. Participants will consider the problem of a user trying to find an app on a mobile phone or tablet. The goal here is to find the app that the user wants to use. So the focus will be on building a cost function to model the different ways in which the user can find the app, e.g. search or browse. To add realism to the scenario, we will consider different types of users, ones that can remember where the app is, and those who cannot (i.e., best and average/worse cases) as well as consider how screen size and app icon size can be modelled to arrive at an optimal size and layout for such interaction. During the practical session, participants will be encouraged to abstract away the details to form a representation of the problem, and identify the main variables that are likely to influence the interaction (i.e., number of apps on the phone, the number apps per page, the cost of moving between screens, etc). These will be used to formulate the costs, then we will be able to reason about when it is better to search for an app and when it is better to browse. We will also consider alternative designs, such as presenting the most used apps first, or a hierarchical browsing structure, and whether they are likely to be more efficient or not, or under what circumstances.

2. INTENDED LEARNING OUTCOMES

By the end of the tutorial, participants should be able to:

- Define and describe the different types of models
- Explain the rationale for formal models of ISR
- Describe and define the components and optimisation model
- Describe and explain the main contemporary models
- Explain and infer the predicted user behaviour given these contemporary models
- Design a formal model and generate hypotheses regarding user behavior

3. BIOGRAPHY

Leif Azzopardi is a Senior Lecturer within the School of Computing Science at the University of Glasgow. His research focuses on building formal models for Information Retrieval - usually drawing upon different disciplines for inspiration, such as Quantum Mechanics, Operations Research, Microeconomics, Transportation Planning and Gamification. Central to his research is the theoretical development of formal models for Information Retrieval, where his research interests include:

- Statistical Language Models for the retrieval of documents, sentences, experts and other information objects [12, 24];
- Probabilistic models of user interaction and the simulation of users for evaluation [2, 8, 9];
- Microeconomic models of information interaction, specifically how cost and effort affect interaction and performance with search systems [3];
- Methods which assess the impact of search technology on society in application areas such as search engine bias and the accessibility of e-Government information [11], and;
- Search for fun (i.e. the SINS of users) [4].

He received his Ph.D. in Computing Science from the University of Paisley in 2006, and he received a First Class Honours Degree in Information Science from the University of Newcastle, Australia, 2001. In 2010, he received a Post-Graduate Certificate in Academic Practice and has been lecturing at the University of Glasgow since then. He has given numerous invited talks on Formal Models of Information Seeking and Retrieval throughout the world and lectured at the Information Foraging Summer School (2011, 2012 and 2013) and Symposium of Future Directions in Information Access (2007-2013). His work on economic models [7] for search received a best paper honourable mention at SIGIR 2014. He recently released a free online book called, *How to Tango with Django* which is a noob's guide to web development in Python's Django (available free at: `www.tangowithdjango.com` [10]).

Guido Zuccon is a Lecturer within the School of Information Systems at the Queensland University of Technology. His research interests include formal models of search, ranking principles for information retrieval, and retrieval models for health search. Guido has actively contributed to the area of document ranking and search result diversification. During his PhD he performed an extensive analysis of document ranking principles [49] and introduced the quantum probability ranking principle [51, 47] and was the first to empirically evaluate the interactive PRP [48]. His work on formal models of search result diversification based on facility location analysis [50] received the best paper award at ECIR 2013 and then in 2014 he received a best reviewer award at ECIR.

He received a Ph.D. in Computing Science from the University of Glasgow in 2012, and he received a Master in Computer Engineering with summa cum laude from the University of Padua, Italy, in 2007. Before joining the Queensland University of Technology as a lecturer in 2014, he was a postdoctoral research fellow at the CSIRO, Australia.

4. REFERENCES

[1] C. W. Axelrod. The economic evaluation of information storage and retrieval systems. *Information Processing & Management*, 13(2):117–124, 1977.

[2] L. Azzopardi. Query side evaluation: an empirical analysis of effectiveness and effort. In *Proc. of SIGIR*, pages 556–563. ACM, 2009.

[3] L. Azzopardi. The economics in interactive information retrieval. In *Proc. of SIGIR*, pages 15–24. ACM, 2011.

[4] L. Azzopardi. Searching for unlawful carnal knowledge. In *Proc. of SIGIR Workshop: Search for Fun*, volume 11, pages 17–18, 2011.

[5] L. Azzopardi. Economic models of search. In *Proceedings of the 18th Australasian Document Computing Symposium*, ADCS '13, pages 1–1, 2013.

[6] L. Azzopardi. Economic models of search. In *Proceedings of the 18th Australasian Document Computing Symposium*, ADCS '13, pages 1–1, 2013.

[7] L. Azzopardi. Modelling interaction with economic models of search. In *Proc. of SIGIR*, pages 3–12, 2014.

[8] L. Azzopardi and M. de Rijke. Automatic construction of known-item finding test beds. In *Proc. of SIGIR*, pages 603–604, 2006.

[9] L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six european languages. In *Proc. of SIGIR*, pages 455–462. ACM, 2007.

[10] L. Azzopardi and D. Maxwell. Tango with django: a begginners guide to web development in python / django, 2013.

[11] L. Azzopardi and V. Vinay. Retrievability: An evaluation measure for higher order information access tasks. In *Proc. of CIKM*, pages 561–570, 2008.

[12] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proc. of SIGIR*, pages 43–50, 2006.

[13] M. J. Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Information Review*, 13(5):407–424, 1989.

[14] M. J. Bates. Training and education for online. chapter Information search tactics, pages 96–105. Taylor Graham Publishing, London, UK, 1989.

[15] N. J. Belkin. Some(what) grand challenges for information retrieval. *SIGIR Forum*, 42:47–54, 2008.

[16] N. J. Belkin. Some(what) grand challenges for information retrieval. *SIGIR Forum*, 42(1):47–54, 2008.

[17] N. J. Belkin, R. N. Oddy, and H. M. Brooks. Ask for information retrieval: part i: background and theory; part ii: results of a design study. *Journal of Documentation*, 38(2) 61-71 and 38(3) 145-164, 1982.

[18] U. Birchler and M. Butler. *Information economics.* Routledge, 2007.

[19] E. L. Charnov. Optimal foraging: attack strategy of a mantid. *The American Naturalist*, 110(971):141–151, 1976.

[20] E. H. Chi, P. Pirolli, and J. Pitkow. The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a web site. In *Proc. of SIGCHI*, pages 161–168. ACM, 2000.

[21] E. Coiera. Information economics and the internet. *Journal of the American Medical Informatics Association*, 7:215–221, 2000.

[22] M. D. Cooper. A cost model for evaluating information retrieval systems. *Journal of the American Society for Information Science*, pages 306–312, 1972.

[23] S. Erdelez. Information encountering: a conceptual framework for accidental information discovery. In *Proc. of ISIC*, pages 412–421, 1997.

[24] R. T. Fernández, D. E. Losada, and L. A. Azzopardi. Extending the language modeling framework for sentence retrieval to include local context. *Information Retrieval*, 14(4):355–389, 2011.

[25] N. Fuhr. A probability ranking principle for interactive information retrieval. *Information Retrieval*, 11(3):251–265, 2008.

[26] F. S. Hillier and G. J. Lieberman. Introduction to operations research. *NY, US*, 2001.

[27] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context.* Springer-Verlag New York, Inc., 2005.

[28] K. Järvelin. IR research: systems, interaction, evaluation and theories. *SIGIR Forum*, 45(2):17–31, 2012.

[29] K. Järvelin and T. D. Wilson. On conceptual models for information seeking and retrieval research. *Information Research*, 9(1):9–1, 2003.

[30] C. C. Kuhlthau. Developing a model of the library search process: Cognitive and affective aspects. *RQ*, pages 232–242, 1988.

[31] K. G. Murty. Optimization models for decision making: Volume. *University of Michigan, Ann Arbor*, 2003.

[32] P. Pirolli and S. Card. Information foraging. *Psychological Review*, 106:643–675, 1999.

[33] A. J. Repo. The value of information: Approaches in economics, accounting, and management science. *Journal of the American Society for Information Science*, 40(2):68–85, 1989.

[34] H. L. Resnikoff, H. Resenikoff, and H. Resnikoff. *The illusion of reality.* Springer-Verlag New York, 1989.

[35] S. E. Robertson. The probability ranking principle in ir. *Journal of Documentation*, 33(4):294–304, 1977.

[36] D. H. Rothenberg. An efficiency model and a performance function for an IR system. *Information Storage and Retrieval*, 5(3):109 – 122, 1969.

[37] D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card. The cost structure of sensemaking. In *Proc. of INTERACT/SIGCHI*, pages 269–276, 1993.

[38] P. E. Sandstrom. An optimal foraging approach to information seeking and use. *The library quarterly*, pages 414–449, 1994.

[39] H. A. Simon. A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118, 1955.

[40] H. A. Simon. Theories of bounded rationality. *Decision and organization*, 1:161–176, 1972.

[41] D. Stephens and J. Krebs. Foraging theory. *Princeton: Princeton University Press*, 1(10):100, 1986.

[42] G. J. Stigler. The economics of information. *The journal of political economy*, 69(3):213–225, 1961.

[43] H. R. Varian. How to build an economic model in your spare time. *American Economist*, 41:3–10, 1997.

[44] H. R. Varian. Economics and search. *SIGIR Forum*, 33(1):1–5, 1999.

[45] M. L. Weitzman. Optimal search for the best alternative. *Econometrica: Journal of the Econometric Society*, pages 641–654, 1979.

[46] T. D. Wilson. Human information behavior. *Informing science*, 3(2):49–56, 2000.

[47] G. Zuccon and L. Azzopardi. Using the quantum probability ranking principle to rank interdependent documents. In *Proc. of ECIR*, pages 357–369. Springer, 2010.

[48] G. Zuccon, L. Azzopardi, and C. van Rijsbergen. The interactive prp for diversifying document rankings. In *Proc. of SIGIR*, pages 1227–1228. ACM, 2011.

[49] G. Zuccon, L. Azzopardi, and C. K. Van Rijsbergen. An analysis of ranking principles and retrieval strategies. In *Proc. of ICTIR*, pages 151–163. Springer, 2011.

[50] G. Zuccon, L. Azzopardi, D. Zhang, and J. Wang. Top-k retrieval using facility location analysis. In *Proc. of ECIR*, pages 305–316. 2012.

[51] G. Zuccon, L. A. Azzopardi, and K. Rijsbergen. The quantum probability ranking principle for information retrieval. In *Proc. of ICTIR*, pages 232–240, 2009.