

Understanding and Improving Search Using Large-Scale Behavioural Data

An Overview of the 2015 Tony Kent Strix Award Annual Lecture

Gary Horrocks, Editor eLucidate

gary.horrocks@gmail.com

On the 6th November 2015 the winner of the 2014 Tony Kent Strix Award presented the Strix annual lecture at The Geological Society in London. Susan Dumais (Microsoft Research) delivered a fascinating and thought-provoking presentation on ‘Understanding and Improving Search Using Large-Scale Behavioural Data.’ I’d like to provide my synopsis and interpretation here, and encourage you to view the [full video](#).

The rise of web-based search systems over the past decade has enabled information scientists to develop powerful large-scale behavioural logging technologies that provide a unique insight into ‘what searchers do’; how people interact with web-based search systems. This ability to gather traces of human behaviour on an extensive scale and speed provides the backdrop to innovation and improvement in search, complementing other forms of experimental research that observes how people engage with search systems including, for example, controlled lab-based observational studies.

It’s hard to believe how far search has advanced in a mere two decades, from the early days of NCSA’s Mosaic browser in 1995. Today web search is pervasive; what Dumais describes as ‘the core fabric of people’s lives.’ Information retrieval, in turn, is no longer monopolised by a minority, but has ‘transformed from an arcane skill that was possessed by information scientists’ to a daily routine. Dumais provides some breath-taking statistics on the state of play in 2015: a billion web sites, trillions of pages indexed by search engines, billions of web searches per day, multiple modes of access.

If search was still in its infancy twenty years ago, behavioural logging was barely conceived. We had no real idea what people were searching for, nor any insight into what they were doing or how they were interacting with search engines. It was difficult back then to optimise search systems or provide a better user experience without the evidential data about online behaviours.

Dumais demystified observational logs by making an analogy with print and how people interact with physical books. Invariably a book will fall open on a specific page if a chapter has been heavily read. Pages corners are often folded, highlighter pens used to identify specific memorable content, annotations made in the margins. We can trace human interaction with a book using these basic means. Observational logs work in a similar context, capturing online interactions. What was the original search query? Was it

reformulated with alternative terminology? What were the results? What was the page scrolling behaviour? What results were clicked on? What was the dwell time on any specific web page?

While lab-based observational studies are carefully controlled with explicit tasks, the capture of large-scale behavioural log data in situ has immense implicit and complementary benefits. They are:

- In real time: there is immediate access to trends, breaking news, reactions to world events as they happen
- Real world: with nobody observing you or giving you controlled tasks they capture warts and all behaviour (You are not likely to search for pornography in a controlled lab-based study!)
- Large-scale (Researchers have access to millions of searches, many unique)
- Diverse: there is access to a multiplicity of behaviours and motivations for searching

Observational logging has huge potential, providing unique insight into the complex world of typology and query formulation. How are people articulating what they are looking for? How do they use (or misuse) terminology? What are the patterns of misspellings? How do they disambiguate synonymous terms or formulate complex questions? Is their query syntax basic or are they utilising advanced search functionality? How do they respond to search results, navigate lists and link to other web pages?

The data captured by the logging of behavioural interactions is critical to the quality of modern web search. The information can be used to:

- Improve system performance
- Improve ranking algorithms
- Enable spelling correction and auto-completion
- Support query suggestion and reformulation
- Improve the presentation of results
- Improve how people interact with the system

This type of research also has the immense potential to identify larger scale societal issues, for example, improving the speed and scale of detection of a medical trend that might save lives. We are all aware of the Doctor Google phenomenon, that health, medical and drug information is a major motive for search (alongside sex, shopping, leisure, travel and news.) Dumais provided a powerful example of the use of search logs to extract data on an adverse drug effect and drug interaction. A 2011 report flagged up that two key drugs Paroxetine and Pravastatin, for depression and lowering cholesterol respectively, were leading to incidences of hyperglycaemia. Pre-2011 search logs that featured the two drugs in combination were analysed and identified a prevalence of terms related to the condition: ‘thirsty’, ‘frequent urination’ and ‘high blood sugar’, for example.

There are obvious drawbacks to observational logs, specifically the sheer noise and abundance of data that has to be sifted and made sense of. Logs can convey what people

are doing, but not why they are doing it. What is the information need and motivation behind the search? Were the results satisfactory, and did they fill a knowledge gap? Is a quick exit after a search a sign of user satisfaction or exasperated abandonment? Whatever the answers to these complex questions, large-scale observational logs complement alternative information retrieval research techniques, and have immense potential for information science and the design, evaluation and evolution of search technologies.

<http://research.microsoft.com/~sdumais>