# Neural Methods for Effective, Efficient, and Exposure-Aware Information Retrieval

Bhaskar Mitra
University College London
*bhaskar.mitra.15@ucl.ac.uk*

**Abstract**

Neural networks with deep architectures have demonstrated significant performance improvements in computer vision, speech recognition, and natural language processing. The challenges in information retrieval (IR), however, are different from these other application areas. A common form of IR involves ranking of documents—or short passages—in response to keyword-based queries. *Effective* IR systems must deal with query-document vocabulary mismatch problem, by modeling relationships between different query and document terms and how they indicate relevance. Models should also consider lexical matches when the query contains rare terms—such as a person's name or a product model number—not seen during training, and to avoid retrieving semantically related but irrelevant results. In many real-life IR tasks, the retrieval involves extremely large collections—such as the document index of a commercial Web search engine—containing billions of documents. *Efficient* IR methods should take advantage of specialized IR data structures, such as inverted index, to efficiently retrieve from large collections. Given an information need, the IR system also mediates how much exposure an information artifact receives by deciding whether it should be displayed, and where it should be positioned, among other results. *Exposure-aware* IR systems may optimize for additional objectives, besides relevance, such as parity of exposure for retrieved items and content publishers.

In this thesis, we present novel neural architectures and methods motivated by the specific needs and challenges of IR tasks. We ground our contributions with a detailed survey of the growing body of neural IR literature [Mitra and Craswell, 2018]. Our key contribution towards improving the *effectiveness* of deep ranking models is developing the Duet principle [Mitra et al., 2017] which emphasizes the importance of incorporating evidence based on both patterns of exact term matches and similarities between learned latent representations of query and document. To *efficiently* retrieve from large collections, we develop a framework to incorporate query term independence [Mitra et al., 2019] into any arbitrary deep model that enables large-scale precomputation and the use of inverted index for fast retrieval. In the context of stochastic ranking, we further develop optimization strategies for exposure-based objectives [Diaz et al., 2020]. Finally, this dissertation also summarizes our contributions towards benchmarking neural IR models in the presence of large training datasets [Craswell et al., 2019] and explores the application of neural methods to other IR tasks, such as query auto-completion.

**Awarded by:** University College London, London, UK **on** 28 April 2021.
**Supervised by:** Emine Yilmaz.
**Available at:** https://discovery.ucl.ac.uk/id/eprint/10125532.

## Selected Publications


Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. Overview of the trec 2019 deep learning track. In *Proc. TREC*, 2019. https://trec.nist.gov/pubs/trec28/papers/OVERVIEW.DL.pdf.

Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. Evaluating stochastic rankings with expected exposure. In *Proc. CIKM*, 2020. https://dl.acm.org/doi/10.1145/3340531.3411962.

Bhaskar Mitra and Nick Craswell. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 2018. https://www.nowpublishers.com/article/Details/INR-061.

Bhaskar Mitra, Fernando Diaz, and Nick Craswell. Learning to match using local and distributed representations of text for web search. In *Proc. WWW*, pages 1291–1299, 2017. https://dl.acm.org/doi/10.1145/3038912.3052579.

Bhaskar Mitra, Corby Rosset, David Hawking, Nick Craswell, Fernando Diaz, and Emine Yilmaz. Incorporating query term independence assumption for efficient retrieval and ranking using deep neural networks. *arXiv preprint arXiv:1907.03693*, 2019. https://arxiv.org/abs/1907.03693.