The role of ML in IR research

Hugo Zaragoza
Yahoo Espagne
hugoz@es.yahoo-inc.com

The commercial success of the web is rapidly changing the nature of Information Retrieval (IR) research. Both information (the data) and retrieval (the task) are growing in complexity. As the size of the WWW and corporate intranets grows, the amount of data available becomes larger, but also richer: metadata, link connectivity, and user behaviour are becoming as important as the actual textual content of documents. At the same time, as more and more people begin to use this data, the demand increases for new modes of interaction, new tasks, and new forms of access.

Because of this rapid increase in complexity, Machine Learning (ML) has become an important tool in IR research. ML permeates almost every aspect of IR research today: ranking, categorisation, evaluation, personalisation, question answering, usage mining, monetisation, spam and fraud detection, etc. The fit however is far from perfect: ultimately, IR needs to deal with natural language and human interaction, both of which demand very rich models of representation and inference far beyond current capabilities of ML.

In my talk I will discuss these issues, drawing from examples of my own work, and I will point out some open problems which I feel are most important to the further development of IR research and its commercial applications.