Data-Driven Information Retrieval

Maristella Agosti
University of Padua, Italy
maristella.agosti@unipd.it

Omar Alonso
Microsoft, USA
omar.alonso@microsoft.com

Maarten de Rijke
University of Amsterdam, The Netherlands
derijke@uva.nl

Raffaele Perego
ISTI-CNR, Pisa, Italy
raffaele.perego@isti.cnr.it

1 Background and Scope of the Panel

This paper reports on the *Data-Driven Information Retrieval* panel that was organized at the ECIR 2016 Conference on Monday 21st March 2016 and that took place in the “Sala delle Colonne” of the Botanical Garden of the University of Padua, Italy [6, 7].[1]

The panel originated from the consideration that IR has mostly been concerned with finding the “needle in a haystack” to retrieve the most relevant information able to best address user information needs from huge amounts of data.

The huge amounts of data that we are used to dealing with are chosen, processed, produced and managed directly by IR experts using the methods and the systems that have been built up over the years in the IR area. These include huge data indexes that need to be used to effectively answer user queries. This type of approach is peculiar to the area of IR. Indeed, one striking example is the area of databases where the data to be managed are structured data inserted by users and constitute a well defined set.

The IR area is used to process and manage huge amounts of data that are also rapidly changing over time and that are produced and prepared in an unforeseeable manner. IR processes and represents a huge amount of data through indexes that are a relevant example of big data. Over the decades, the area of IR has been able to react to at least three main changes: 1) the creation of realistic test collections similar in content and size and to be used to test the effectiveness of the IR processes, 2) the move from homogeneous to heterogeneous collections of data, and 3) the need to invent methods and processes also able to manage collections of data derived from scientific experiments that produce streams of massive data over short periods of time.

Nowadays we can observe that in many other related research fields – e.g., machine learning, crowdsourcing, user interaction analysis – there has been a radical change in the approach of extracting and deriving useful information, with a shift towards a more data-driven approach. Can this approach also be useful in the IR area? Are some of the methods we have invented in IR similar to what is happening in those research fields related to the IR area?

[1] `http://www.ortobotanicopd.it/en`

The goals of the panel were to discuss the emergent trends in the area – advantages, pitfalls, implications – to try to give initial directions for the future. In the following the initial position statements made by each panelist are reported together with the discussion that took place. Some of the directions that emerged and that seem necessary to follow are reported.

2 The Right Information to the Right People in the Right Way

For many years, IR focused on "the provision of relevant documents" as the goal of its most fundamental algorithms [2]. This is an outdated perspective. IR systems are best thought of as agents that aim to take the best possible action so as to get the right information to the right people in the right way. Actions range from understanding a user's query and selecting appropriate information sources to ranking results from those sources and producing an effective results page. Each action may influence both the user and the IR system's future state. Success is often measured by interpreting behavioral signals that may be noisy and biased.

In recent years, two essential factors have changed to help us realize the vision of IR systems as the intelligent agents described above. One is the increasing availability of very large volumes of fine-grained interaction data: data that captures, at some level of abstraction, people's information needs and their responses to results such as selecting a query completion, clicking on an item on a search engine result page, re-visiting a result visited during an earlier session, etc. The other is our improved ability to interpret such interaction data with limited or even absent supervision.

To make matters concrete, let's quickly list some examples. We are now able to have IR systems predict query difficulty with reasonable accuracy and without supervision [8]. Further down the stack, advances in click models [4] now allow us to identify different forms of bias in users' interactions so that we are better able to identify the real attractiveness of documents; while early approaches required the manual design of probabilistic graphical models for this purpose, very recent proposals allow us to achieve the same from raw interaction data as input [3], i.e., without having to manually design a model. Interleaving methods allow us to reliably infer users' preferences from interaction data [9]. Ranking documents using their historical click-through rate can improve relevance for frequently occurring queries, i.e., so-called head queries. Until recently it was difficult to use such click signals on non-head queries as they receive fewer clicks. However, with a modest amount of exploration, it is now possible to obtain substantial improvements over a production ranker in terms of page-level online metrics even for so-called torso queries [11].

This upbeat story on the happy marriage of information retrieval and data hungry artificial intelligence methods has a flip side. We are increasingly handing over control to black box algorithms that experiment with us, human users, to be able to learn from our behavior and improve the quality of the results that they produce. The positivistic "IR meets AI" agenda cannot do without progress on a second research agenda. This second agenda is filled with complex issues at the interface where our technology meets society, ranging from "search engines that you can trust" and "search engines that can explain their results" to "search engines that do not experiment more with people than they need to." This second agenda is at least as challenging as the first one and can best be summarized as "responsible information retrieval." The series of workshops on privacy-aware/privacy-preserving information retrieval [10] is a great initiative but we should realize that responsibility in IR is so much more than privacy. The IR community should find algorithmic inspirations in societal concerns and in emerging rules such as the recently adopted EU law whereby those affected by algorithmic decisions are entitled to an explanation of those decisions [5]: these concerns and rules should not block algorithmic progress but they should motivate us to design IR algorithms in which the concerns are addressed by design.

3 The Data Stack in Information Retrieval

The amount of data that is crawled, indexed, and retrieved for presenting relevant results to users in a commercial search engine is massive. Also, equally massive data sets are used internally by service providers for running experiments and testing improvements with the end goal of enhancing user satisfaction. As we see data science and data intensive applications play a pivotal role in the real world, there is an opportunity to look at IR from a data perspective. This alternative view is complementary to the traditional user-centric notion that we are all familiar with in IR. The database community has identified big data as a challenge for the new generation of database infrastructure [1]. The IR community is also at the core of the big data revolution and needs to identify similar challenges.

An IR system has many different data layers that are available and practitioners and researchers should be familiar with different techniques. The proposed data stack consists of the following parts:

- Ingestion and processing of raw data. Examples: crawling, data cleaning, near-duplicate detection, etc.
- Annotations for augmenting the ingested data and adding, if possible, semantic information and other metadata. Examples: named-entity detection, information extraction, content classification, etc.
- Indexing and ranking of high quality content. Examples: efficient data structures, relevance ranking, etc.
- Behavioral data for capturing user activity via search query logs, clicks, link sharing, etc.
- Experimentation and analysis infrastructure for evaluation, exploration, and exploitation.

The idea behind the data stack is to think about IR challenges and opportunities by focusing on the many data aspects that are available and the type of insights that we can gain by exploring and analyzing such data. New scenarios and problems can be identified by looking at data as it changes and evolves over time. This is independent of the data size, that is, small data (e.g., personal email, calendar, etc.) or large data (e.g., the web, a social network, etc.). Thinking in terms of a data stack helps in research and development as every improvement benefits end users.

4 Discussion and Future Directions

The availability of massive datasets and the advances in machine learning and high performance computing have enabled a radical change in IR during the last few years. The change is firstly methodological. IR was traditionally a data-informed, problem-driven discipline where researchers designed suitable methods for their problems. It progressively moved to a data-driven discipline where the starting point is not a (analytical or statistical) model to be fitted on the data, but rather a rich data set providing opportunities for better modeling the problem or gaining new insights into it. Mandatory skills for today's IR researchers and engineers encompass the whole scalable data stack discussed above and the analytical ability to manage and interpret large data sources in order to rapidly prototype and experiment new solutions.

As a direct result of this transformation we have now highly scalable solutions for tasks such as image understanding and semantic similarity that were previously too complex to be handled effectively by IR systems. In addition, the state-of-the-art solutions to traditional IR tasks such as ranking, recommendation, user modeling, document and query understanding, exploit machine learning and attain the best performance when trained with large amounts of data.

All that glitters is not gold, and we must be conscious that large data and the knowledge gained from them are not objective "per se" since data production is in any case human-mediated. Hidden biases in both the collection and analysis of data are just around the corner, and they ask for bringing a higher level of attention and awareness to our research. On the other hand, data-driven IR stimulates challenging opportunities for actionable and high-impact research.

The ECIR panel and the follow-up discussion touched on a number of important points deserving further attention by our community:

- Semi-supervised and data augmentation techniques making efficient reuse of available data are increasingly important. The availability of data in some domains will always be scarce, and in any case large corpora of human-labeled data are too expensive to produce;

- Awareness of where the data comes from and how it was gathered is a key factor to mitigate cognitive biases or ethical issues that might influence the interpretation of data and the derived models;

- The size and complexity of data ask for a robust data stack and novel algorithmic solutions. The challenging goal here is to make all steps fast and scalable, from data gathering and cleaning up to validation, deployment and use of the extracted models in large-scale applications;

- Black-box modeling is the prevailing approach so far, but there is an increasing demand for transparent models that can be inspected and explain their behavior.

Acknowledgments

We would like to thank all the participants in the panel for their contributions to the discussion and to building the context of a possible new area of Information Retrieval.

References

[1] D. Abadi, R. Agrawal, A. Ailamaki, M. Balazinska, P. A. Bernstein, M. J. Carey, S. Chaudhuri, J. Dean, A. Doan, M. J. Franklin, J. Gehrke, L. M. Haas, A. Y. Halevy, J. M. Heller-

stein, Y. E. Ioannidis, H. V. Jagadish, D. Kossmann, S. Madden, S. Mehrotra, T. Milo, J. F. Naughton, R. Ramakrishnan, V. Markl, C. Olston, B. C. Ooi, C. Ré, D. Suciu, M. Stonebraker, T. Walter, and J. Widom. The Beckman report on database research. *Commun. ACM*, 59(2):92–99, February 2016.

[2] N. Belkin. People interacting with information. *SIGIR Forum*, 49(2):13–27, December 2015.

[3] A. Borisov, I. Markov, M. de Rijke, and P. Serdyukov. A neural click model for web search. In *WWW 2016: 25th International World Wide Web Conference*, pages 531–541. ACM, April 2016.

[4] A. Chuklin, I. Markov, and M. de Rijke. *Click Models for Web Search*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, August 2015.

[5] EU. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, 59, 2016.

[6] N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. Di Nunzio, C. Hauff, and G. Silvello. Preface. In N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. Di Nunzio, C. Hauff, and G. Silvello, editors, *Advances in Information Retrieval. Proceedings of the 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016*, pages v–vii. Springer, April 2016.

[7] N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, J. Kekäläinen, P. Rosso, P. Clough, G. Pasi, C. Lioma, S. Mizzaro, G. Di Nunzio, C. Hauff, O. Alonso, P. Serdyukov, and G. Silvello. Report on ECIR 2016: 38th European Conference on Information Retrieval. *SIGIR Forum*, 50(1):12–27, June 2016.

[8] J. He, M. Larson, and M. de Rijke. Using coherence-based measures to predict query difficulty. In *30th European Conference on Information Retrieval (ECIR 2008)*, number 4956 in LNCS, pages 689–694. Springer, April 2008.

[9] K. Hofmann, S. Whiteson, and M. de Rijke. A probabilistic method for inferring preferences from clicks. In *20th ACM Conference on Information and Knowledge Management (CIKM 2011)*, pages 249–258. ACM, October 2011.

[10] Privacy Preserving IR workshop series. URL: `https://privacypreservingir.org`, 2016.

[11] M. Zoghi, T. Tunys, L. Li, D. Jose, J. Chen, C. M. Chin, and M. de Rijke. Click-based hot fixes for underperforming torso queries. In *SIGIR 2016: 39th international ACM SIGIR conference on Research and development in information retrieval*, pages 195–204. ACM, July 2016.