Large-Scale Information Retrieval Experimentation with Terrier

Rodrygo L. T. Santos
School of Computing Science
University of Glasgow
G12 8QQ Glasgow, UK
rodrygo@dcs.gla.ac.uk

Richard McCreadie
School of Computing Science
University of Glasgow
G12 8QQ Glasgow, UK
richardm@dcs.gla.ac.uk

Vassilis Plachouras
PRESANS, X-TEC
École Polytechnique
91128 Palaiseau, France
vplachouras@acm.org

ABSTRACT

This tutorial aims to provide a practical introduction to conducting large-scale information retrieval (IR) experiments, using Terrier[1] as an experimentation platform. Written in Java, Terrier provides an open-source, feature-rich, flexible, and robust environment for large-scale IR experimentation. This tutorial will cover the experimentation process end-to-end, from configuring Terrier to a particular experimental setting, to efficiently indexing a document corpus and retrieving from it, and to evaluating the outcome. Moreover, it will describe how to use and extend the platform to one's own needs, and will be illustrated by practical research-driven examples. As a half-day tutorial, it will be split into two major sessions, with each session comprising both background information and practical demonstrations. In the first session, we will provide an overview of several aspects of large-scale IR experimentation, spanning areas such as indexing, data structures, query languages, and advanced retrieval models, and how these are implemented within Terrier. In the second session, we will discuss how to extend Terrier to conduct one's own experiments in a large-scale setting, including how to facilitate the evaluation of non-standard IR tasks through crowdsourcing. The practical demonstrations will cover recent use cases identified from Terrier's online discussion forum, so as to provide attendees with concrete examples of what can be done within Terrier.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process*

General Terms

Algorithms, Experimentation, Performance

Keywords

Terrier, large-scale experimentation, information retrieval

[1] `http://terrier.org`

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
ACM 978-1-4503-0717-8/11/10.

Presenters' Biography

Rodrygo Santos is a PhD student at the School of Computing Science of the University of Glasgow. He holds a BSc (2005) and an MSc (2007) degrees from the Dept. of Computer Science of the Federal University of Minas Gerais, Brazil. As a member of the Terrier Team since 2008, he has extensive experience in Terrier, extending it to a variety of application domains. These include blog search, opinion retrieval, expert search, entity search, Web mining, search result diversification, and learning-to-rank. Rodrygo has published several papers in major IR conferences and journals using Terrier, including SIGIR, WWW, CIKM, ECIR, RIAO, ICTIR, IP&M and the IR Journal, as well as international evaluation forums, such as TREC and NTCIR.

Richard McCreadie is a PhD student at the School of Computing Science of the University of Glasgow. He holds a BSc (Hons) from the University of Glasgow (2008) and is a member of the Terrier Team since 2008. In particular, he has developed the MapReduce indexing functionality for massive-scale IR that Terrier currently supports. His work using Terrier has been published in several major IR conferences and journals, such as SIGIR, RIAO, and IP&M, as well as on the TREC evaluation forum. Furthermore, Richard is also an expert in crowdsourcing solutions to common IR problems. He has presented new crowdsourcing approaches at the influential crowdsourcing workshops CSE 2010 (SIGIR) and CSDM 2011 (WSDM). Moreover, he designed, implemented and validated the first successful instance of crowdsourcing relevance assessments for TREC during 2010.

Dr. Vassilis Plachouras is a researcher at PRESANS, a start-up company from École Polytechnique (France). He holds a BEng degree in Electrical and Computer Engineering from the National Technical University of Athens (2001), a PhD in IR from the University of Glasgow (2006), and has worked as a post-doctoral researcher on Web search at Yahoo! Research, Spain. His research interests include Web IR, architectures for large-scale Web search engines, online advertising, and the applications of machine learning for ranking and classification. His publications include over 30 articles in peer-reviewed journals and international conferences, and he has received the best-paper award in CIKM 2009. He has served as a PC member of major international conferences and journals, and as the workshops and tutorials PC chair of ECIR 2008. Moreover, he organised the Workshop for Large-Scale and Distributed Systems for IR (LSDS-IR'07) in SIGIR 2007. He has been involved in the design and development of Terrier since its inception.