Knowledge Lab

Birkbeck UNIVERSITY OF LONDON

Research Student
Adam Foster

Supervisors
Mark Levene
Dell Zhang

Birkbeck UNIVERSITY OF LONDON

Department of Computer Science and Information Systems

Predictive Analytics at Scale

Research Aims

In almost every industry today there is an explosion of information, both structured and unstructured, created by computing systems, sensors and ubiquitous devices everywhere. The massive scale of these “Big Data” have led to a new class of techniques for analysis, in-order to gain knowledge and insight. Our research focuses on a novel type of machine learning based on Predictive Suffix Trees. Traditionally these have not been applied at scale due to their relatively poor performance, however with the use of modern partitioning build algorithms and supporting data-structures it is possible to build indexes on datasets and data-types previously unattainable to the Suffix Tree structure.

Research Methodology

Prior research has focused primarily on build efficiency however our research focus here is on scalability, thus build time and on-disc size are sacrificed in favour of parallelization and memory efficiency. In the first instance, the Top Down Disk build method of Suffix Tree construction was implemented in Python 2.7 and Numpy to give a test bed for further research. This on its own however is not sufficient to use as a predictive data structure. Further work is being done to generalise the implementation to remove the initial dependency on use of the Latin alphabet, as well as supplementing the core data structure with additional meta-data to support prediction. Figure 1 shows the Suffix Tree and on-disc representation for the String: ATTAGTACA$

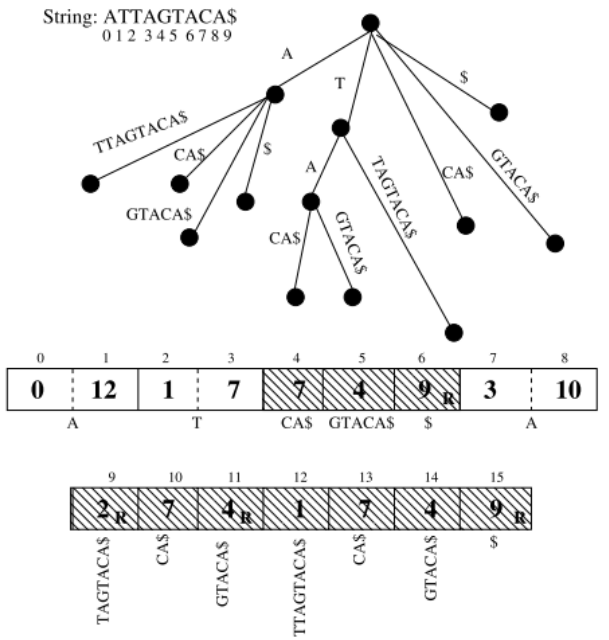

Figure 1: Top Down Disk Suffix Tree Representation

Publications

M. Barsky, U. Stege, and a. Thomo. A survey of practical algorithms for suffix tree construction in external memory. Software - Practice and Experience, 40(11):965–988, 2010.

S Tata, R a Hankins, and J M Patel. Practical suffix tree construction. Data Base, 30:47, 2004.
015)

Nikos Karampatziakis, Nikos Karampatziakis, Dexter Kozen, and Dexter Kozen. Learning prediction suffix trees with Winnow. Proceedings of the 26th Annual International Conference on Machine Learning - ICML ’09, pages 1–8, 2009.

Ron Begleiter, Ran El-Yaniv, and Golan Yona. On Prediction Using Vari- able Order Markov Models. Journal of Artificial Intelligence, 22:385–421, 2004.