The Flamingo Software Package on Approximate String Queries

Chen Li*

Department of Computer Science
UC Irvine, CA 92697, USA
chenli@ics.uci.edu

Abstract. An important operation in data cleaning is similarity search on textual strings. A simple example is "finding actor names similar to `schwarzeneger`," given the fact that few people know the exact spelling of our former governor in California. It is challenging to support this operation efficiently on large amounts of data. Despite its importance, the problem did not receive enough attention in the research community a decade ago. In this talk, I will give an overview of recent results on this problem, and describe the development history of the Flamingo package, an open-source software that supports efficient approximate string queries. I will also describe my outreach activities to apply our research results of data cleaning in real applications, which led to a startup called Bimaple that specializes in powerful instant search on large data sets.

Keywords: Data Cleaning, Flamingo Package, Approximate String Search

* This research is partially supported by the US NSF CAREER award IIS-0238586, the NSF award IIS-0742960, the NSF award IIS-0844574, the NSF award 1030002, the NSF award 0331707, the National Nature Science of China 60828004, a Google Research Award, and a gift fund from Microsoft.