UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

DIPARTIMENTO DI
INFORMATICA, SISTEMISTICA E
COMUNICAZIONE

Viale Sarca, 336
20126 Milano

SEMINAR ANNOUNCEMENT

Monday October 21, 2024

at 02:30 pm

Room "Sala Seminari" - Abacus Building (U14)

Utility-oriented String Mining

Speaker
Giulia Bernardini
University of Trieste

Abstract
A string is often provided with numerical scores (utilities) which quantify the importance, interest, profit, or risk of the letters occurring at every position of the string. For example, every DNA fragment produced by modern sequencing machines comes with a confidence score per position. Motivated by the abundance of strings with utilities, we introduce Utility-oriented String Mining (USM), a natural generalization of the classic frequent substring mining problem. Given a string S of length n and a threshold $\mathcal{V}$, USM asks for every string R whose utility U(R) is at least $\mathcal{V}$, where U is a function that maps R to a utility score based on the utilities of all letters of every occurrence of R in S. In addition, our work makes the following contributions: (1) We identify a class $\mathbb{U}$ of utility functions for which USM admits an $\mathcal{O}(n2)$-time algorithm. (2) We prove that no listing algorithm solves the USM problem in subquadratic time for every utility function, or even for every function in $\mathbb{U}$. (3) We propose an $\mathcal{O}(n \log n)$-time algorithm that solves USM for a class of monotone functions from $\mathbb{U}$. (4) We design another $\mathcal{O}(n \log n)$-time algorithm for the same problem that is comparable in runtime but offers drastic space savings in practice when, in addition, a lower bound on the length of the output strings is provided as input. (5) We demonstrate experimentally using publicly available, billion-letter datasets that our algorithms are many times more efficient, in terms of runtime and/or space, compared to an Apriori-like baseline which employs advanced string processing tools.

Link to the publication: https://doi.org/10.1137/1.9781611978032.22

Short Bio:
Giulia is an assistant professor (RTDa) at the University of Trieste. She completed a PhD in Computer Science at the University of Milano-Bicocca in February 2021, supervised by Prof. Paola Bonizzoni and Prof. Nadia Pisanti.

Before joining the University of Trieste, she spent one year as a postdoc at CWI in Amsterdam. Her main research interests are in the field of Combinatorial Algorithmics and Optimisation. She primarily works on problems arising in Computational Biology (analysis of genomes and phylogenies) and Data Mining (data privacy and pattern mining).

contact person for this Seminar: Prof.ssa Paola Bonizzoni (paola.bonizzoni@unimib.it)