M.Sc. “Data and Web Science”
Topics for Diploma Theses 2019

Supervisor: Anastasios Gounaris (gounaria@csd.auth.gr)

Topic 1: Advanced Data Analytics

Description: This thesis topic aims to extend state-of-the-art techniques in a principled and well-defined manner with a view to deriving novel and more efficient solutions. There are five directions that can be followed:

a) Extend the framework in [NG19] with an additional predictive maintenance technique and functionalities (GUI, parameter tuning).

b) Extend the prediction and ensemble techniques in [KWQ18].

c) Extend the scalable clustering technique in [GLT16] with advanced features.

d) Implement, validate and evaluate techniques for scalable single-link hierarchical clustering.

e) Implement a parallel and/or streaming version of the PC causality structure detection technique in [Sha19].

Goal: Develop new techniques

Required Background Knowledge: data mining, at least one of the following programming languages: scala, java, R, python

Comments: can lead to publication

References


[GLT16]: Frank Gouineau, Tom Landry, Thomas Triplet: PatchWork, a scalable density-grid clustering algorithm. SAC 2016: 824-83


Topic 2: Evolving Networks (co-supervised with K. Tsichlas)

Description: In the last years, we have been developing novel techniques for storing dynamic (evolving) graphs [KGT19]. In this topic, the implementation in https://github.com/hinodeauthors/hinode will be extended with new features.

Goal: extend current implementation

Required Background Knowledge: databases, algorithms, java

Comments: can lead to publication

References

**Topic 3: Scalability Evaluation**

**Description:** Scalability is currently offered by several platforms and algorithms, however it is still difficult to be attained in practice due to several reasons, such as parameter tuning. This topic comes into two flavors:

a) Check the scalability of established techniques, such as [http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview](http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview), using a commodity machine.

b) Validate and extend the techniques in [GT18] using the latest Spark version and a cluster consisting of few powerful machines.

**Goal:** derive detailed report on and resolve scalability issues

**Required Background Knowledge:** Java, Scala, Spark, C++

**Comments:** can lead to publication

**References**