Warsaw University of Technology | Doctoral School No. 3

Course offered in the Doctoral School No. 3 – Spring semester of the 2021/2022 academic year

TITLE
Data Mining
CONDUCTING UNIT
Doctoral School No. 3
SCIENTIFIC DISCIPLINE
Information and communication technology
IMPLEMENTING UNIT
103000 - Faculty of Electronics and Information Technology
SUMMARY DESCRIPTION
The contents of the course consists of a wide range of topics within the discipline of data mining. The methods of discovering various knowledge types from large data resources, as well as methods of efficient knowledge acquisition by using concise lossless representations will be presented. Efficient methods of searching for duplicate objects, clustering and classification of data, capable of performing these tasks even several orders of magnitude faster than by using standard algorithms will be depicted. Methods for discovering functional and approximate dependencies between sets of attributes will be also shown. Reasoning from incomplete data and from partial knowledge will be presented. Prerequisites: Knowledge of algorithms and data structures as well as the ability to program in at least one of the following languages: C, C ++, C #, Python.
FULL DESCRIPTION
Lecture contents • Data mining as a multidisciplinary area: Roots and development of data mining area. Current challenges in data mining. Classification of data mining tasks. Data Mining in Knowledge Discovery process. • Frequent patterns and association rules: Scalable methods of discovering frequent patterns and association rules in transactional and relational databases. Modifications of algorithms capable of dealing with hierarchy and negation. Specifying constraints in a data mining language. Usage of imposed constraints for efficient reduction of a discovery process.

- Evaluation measures of association rules: Properties of evaluation measures of association rules such as lift, certainty factor, dependence factor, odds ratio and growth ratio.
- Concise models of frequent patterns: Generators, closed itemsets and generalized-disjunction-free sets as basic elements of lossless representations of frequent patterns. Discovery of concise representations of frequent patterns. Usage of the models for derivation of all frequent patterns.
- Concise models of association rules: Generators and closed itemsets as building blocks of lossless representations of association rules such as representative rules, minimal non-redundant rules and rule templates. Mechanisms of deriving association rules from these representations.
- Other patterns and rules: Methods of discovering other patterns such as sequential patterns and sequential rules, contrast patterns, (rough set) decision rules.
- Similarity and distance measures of objects: Efficient methods of discovering objects that are most similar (or nearest) with respect to the measures such as the Minkowski distance as well as the Jaccard, Tanimoto, cosine and Gower similarity.
- Clustering and noise detection: Density based methods of clustering objects and discovering anomalies such as DBSCAN and NBC and their efficient modifications based on the triangle inequality such as TI-DBSCAN and TI-NBC or based on the VP-tree.
- Classification: Using contrast patterns in classification.
- Functional and approximate dependencies: Scalable methods of discovering functional and approximate dependencies in large databases.
- Reasoning under incompleteness: Legitimate approach to reasoning from data with missing values. Mining from partial knowledge.

Project contents

A project task is to design, implement in C, C++, C# or Python and perform an experimental evaluation of selected data mining algorithms.

LITERATURE

- Han J., Kamber M., Pei J., Data Mining: Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers, 2012
- Kryszkiewicz M., Concise Representations of Frequent Patterns and Association Rules, Prace Naukowe, Elektronika, Oficyna Wydawnicza Politechniki Warszawskiej, z. 142 (2002)
- Ganter B., Wille R., Formal Concept Analysis, Mathematical Foundations, Springer-Verlag, 1999
- a number of recent data mining publications accessible via Internet. The instructor will recommend the respective publications during the course.

LANGUAGE OF THE COURSE	ECTS CREDITS
English	6

TYPE OF CLASSES	NUMBER OF HOURS	COURSE INSTRUCTOR
Lecture	30	Marzena Kryszkiewicz, prof. dr hab. inż.
Project	30	Marzena Kryszkiewicz, prof. dr hab. inż.