Warsaw University of Technology

COURSE OFFERED IN THE DOCTORAL SCHOOL

Code of the course	4606-ES-000000C-0031	Name of the course	Polish	**Metody eksploracji danych w odkrywaniu wiedzy**	
			English	**Data Mining (**EDAMI)	
Type of the course	Special courses				
Course coordinator	prof. dr hab. inż. Marzena Kryszkiewicz				
Implementing unit	WEiTI	Scientific discipline / disciplines*	information and communication technology,		
Level of education	Doctoral studies	Semester	~~Winter and~~ Summer		
Language of the course	English				
Type of assessment:	Graded credit	Number of hours in a semester	60	ECTS credits	5
Minimum number of participants	10	Maximum number of participants	24	Available for students (BSc, MSc)	~~Yes~~/No

Type of classes		Lecture	Auditory classes	Project classes	Laboratory	Seminar
Number of hours	in a week	2	0	2	0	0
	in a semester	15*2=30	0	10*3=30	0	0

* does not apply to the Researcher's Workshop

1. Prerequisites

Knowledge of at least one of the following programming languages: C, C ++, C # or Python

2. Course objectives

The objective of the course is to make students familiar with important topics in the area of data mining. The techniques and algorithms to be presented are of practical value – they are well suited to the discovery of hidden data in real large data sources. As a result of participating in the course, students should be able to discover new, non-trivial and useful knowledge from large data resources, as well as efficiently perform classification, prediction and clustering tasks.

3. Course content (separate for each type of classes)

Lecture

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
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.

4. Learning outcomes

	Learning outcomes description	Reference to the learning outcomes of the WUT DS	Learning outcomes verification methods*
Knowledge			
K01	has knowledge of discovering patterns and dependencies by means of data mining methods	SD_W2	Written test
K02	has knowledge of methods of representing frequent patterns and reasoning about them	SD_W2	Written test
K03	has knowledge of modern data mining technologies	SD_W2	Written test
Skills			
S01	is capable of planning and implementing a knowledge discovery process as well as of interpreting its results	SD_U2, SD_U7	Project evaluation, report evaluation, presentation evaluation; assessment of activity during classes
S02	is capable of presenting a plan, implementation and results of a knowledge discovery process in an oral and written form	SD_U4	Report evaluation, presentation evaluation; assessment of activity during classes
S03	is capable of discovering knowledge using modern data mining technologies	SD_U1	Project evaluation, report evaluation, presentation evaluation; assessment of activity during classes
Social competences			
SC01			

*Allowed learning outcomes verification methods: exam; oral exam; written test; oral test; project evaluation; report evaluation; presentation evaluation; active participation during classes; homework; tests

5. Assessment criteria
In order to pass the course, participants must achieve a pass grade from both course components: the lecture part and the project part. If the course is passed, the final grade is determined as the average of the grades from these two course components.

6. Literature
Basic:
[1] Han J., Kamber M., Pei, J., Data Mining: Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems, 3rd edition, Morgan Kaufmann, 2011
[2] Kryszkiewicz M., Concise Representations of Frequent Patterns and Association Rules, Prace Naukowe, Elektronika, Oficyna Wydawnicza Politechniki Warszawskiej, z. 142, 2002
Additional:
[1] Ganter B., Wille R., Formal Concept Analysis, Mathematical Foundations, Springer-Verlag, 1999
[2] a number of recent data mining publications accessible via Internet. The instructor will recommend the respective publications during the course.

7. PhD student's workload necessary to achieve the learning outcomes**		
No.	Description	Number of hours
1	Hours of scheduled instruction given by the academic teacher in the classroom	30
2	Hours of consultations with the academic teacher, exams, tests, etc.	30
3	Amount of time devoted to the preparation for classes, preparation of presentations, reports, projects, homework	60
4	Amount of time devoted to the preparation for exams, test, assessments	20
	Total number of hours	**140**
	ECTS credits	**5**

** 1 ECTS = 25-30 hours of the PhD students work (2 ECTS = 60 hours; 4 ECTS = 110 hours, etc.)