**Nazwa przedmiotu:**

Metody eksploracji danych w odkrywaniu wiedzy

**Koordynator przedmiotu:**

Marzena Kryszkiewicz

**Status przedmiotu:**

Fakultatywny ograniczonego wyboru

**Poziom kształcenia:**

Studia II stopnia

**Program:**

Informatyka

**Grupa przedmiotów:**

Przedmioty techniczne - zaawansowane

**Kod przedmiotu:**

MED

**Semestr nominalny:**

2 / rok ak. 2018/2019

**Liczba punktów ECTS:**

5

**Liczba godzin pracy studenta związanych z osiągnięciem efektów uczenia się:**

30 godzin wykładu
20 godzin przygotowania do egzaminu
15 godzin ćwiczeń laboratoryjnych
15 godzin przygotowania do ćwiczeń laboratoryjnych
15 godzin spotkań projektowych
30 godzin realizacji zadań projektowych

**Liczba punktów ECTS na zajęciach wymagających bezpośredniego udziału nauczycieli akademickich:**

30 godzin wykładu
15 godzin ćwiczeń laboratoryjnych
15 godzin spotkań projektowych
co daje ok. 2.5 ECTS

**Język prowadzenia zajęć:**

polski

**Liczba punktów ECTS, którą student uzyskuje w ramach zajęć o charakterze praktycznym:**

15 godzin ćwiczeń laboratoryjnych
15 godzin przygotowania do ćwiczeń laboratoryjnych
15 godzin spotkań projektowych
30 godzin realizacji zadań projektowych
co daje ok. 3 ECTS

**Formy zajęć i ich wymiar w semestrze:**

|  |  |
| --- | --- |
| Wykład: | 30h |
| Ćwiczenia: | 0h |
| Laboratorium: | 15h |
| Projekt: | 15h |
| Lekcje komputerowe: | 0h |

**Wymagania wstępne:**

zalecana jest znajomość baz danych

**Limit liczby studentów:**

36

**Cel przedmiotu:**

Celem przedmiotu jest przekazanie studentom wiedzy z dziedziny eksploracji danych. Prezentowane techniki i algorytmy mają istotne znaczenie praktyczne – dobrze nadają się do wydobywania wzorców ukrytych w olbrzymich zasobach danych rzeczywistych. Przewiduje się, że prezentowane metody będą miały pierwszorzędne znaczenie dla ewolucji systemów baz danych do postaci efektywnych systemów baz wiedzy. W rezultacie studenci powinni posiąść umiejętność tworzenia efektywnych systemów odkrywania nowej, nietrywialnej, użytecznej wiedzy z dużych zasobów danych.

**Treści kształcenia:**

Eksploracja danych jako wielodyscyplinarna dziedzina: Przedstawiony będzie rozwój i główne wyzwania w dziedzinie eksploracji danych. Zostanie zaprezentowana szczegółowa klasyfikacja zadań eksploracji danych.
Wstępne przetwarzanie danych dla celów eksploracji: Zostaną przedstawione metody czyszczenia, integracji, transformacji i redukcji (zbędnych) danych oraz użycie hierarchii pojęć do dyskretyzacji.
Język eksploracji danych: Na bazie przykładowego języka eksploracji danych zostanie zaprezentowana koncepcja specyfikowania wiedzy interesującej dla użytkownika.
Techniki budowania deskryptorów pojęć: Zostaną przedstawione techniki opisywania pojęć. Omówiona będzie technika generalizacji atrybutowej oraz jej implementacje. Zostaną przedyskutowane miary statystyczne do celów eksploracji opisowej.
Wzorce częste i reguły asocjacyjne: Zostaną zaprezentowane metody odkrywania wzorców częstych i reguł asocjacyjnych (w tym hierarchicznych) w transakcyjnych i relacyjnych bazach danych. Zostaną opisane strategie wyszukiwania reguł interesujących poprzez efektywne wykorzystanie narzuconych ograniczeń do redukcji przestrzeni rozwiązań.
Zwięzłe reprezentacje wzorców częstych: Zostanie dokonany przegląd zwięzłych bezstratnych reprezentacji wzorców częstych, takich jak: reprezentacje oparte na generatorach, zbiorach zamkniętych, zbiorach (generatorach) wolnych od dysjunkcji. Zostaną przedstawione algorytmy odkrywania tych reprezentacji.
Zwięzłe reprezentacje reguł asocjacyjnych: Dokonany zostanie przegląd zwięzłych bezstratnych reprezentacji reguł asocjacyjnych, w tym reguły reprezentatywne, minimalne reguły nieredundantne, baza podstawowa i informacyjna, baza Duquenne-Guigues'a oraz baza właściwa. Omówione będą mechanizmy wyprowadzania reguł. Przedstawione będą algorytmy bezpośredniego odkrywania reprezentacji reguł asocjacyjnych z reprezentacji zbiorów częstych.
Inne wzorce częste i reguły: Zostaną przedstawione rozszerzenia wzorców częstych i reguł asocjacyjnych na ilościowe reguły asocjacyjne, wzorce sekwencyjne, reguły epizodyczne, częste grafy, klasyfikatory w postaci drzew decyzyjnych i reguł oraz reguły decyzyjne zachowujące przybliżenia dolne i górne zbiorów przybliżonych. Zostaną omówione skalowalne algorytmy odkrywania wyżej wymienionych rodzajów wiedzy.
Grupowanie pojęciowe: Zostanie przedstawiona klasyfikacja metod grupowania pojęciowego. Szczegółowo zostaną omówieni wybrani reprezentanci poszczególnych typów metod grupowania pojęciowego, którzy zapewniają skalowalność rozwiązania.
Wnioskowanie w warunkach niepełności: Zostaną przedstawione zarówno uprawnione mechanizmy wnioskowania z danych z brakującymi wartościami, jak i mechanizmy wyprowadzania wiedzy z dostarczonego jej fragmentu bez dostępu do źródła danych.
Zastosowania eksploracji danych: Omówione będą przykładowe zastosowania eksploracji danych do analizy danych: finansowych, telekomunikacyjnych, biomedycznych i DNA. Zostanie dokonany krótki przegląd wybranych systemów eksploracji danych, które już pojawiły się w komercyjnym użyciu, jak i obiecujących prototypów badawczych.
Zakres laboratorium
Celem laboratorium jest zaznajomienie studentów z nowoczesnymi
technologiami eksploracji danych. W trakcie laboratorium studenci
zapoznają się z możliwościami prowadzenia eksploracji danych w wybranym
komercyjnym systemie, np. IBM Warehouse Design Studio.
Przykładowa tematyka ćwiczeń laboratoryjnych:
Przygotowanie danych do procesu eksploracji: selekcja atrybutów, dyskretyzacja, generowanie danych testowych, ...;
Odkrywanie reguł asocjacyjnych;
Klasyfikacja;
Odkrywanie wzorców sekwencyjnych;
Grupowanie danych.
Zakres projektu
Projekt będzie polegał na zaimplementowaniu i przetestowaniu wybranych
algorytmów odkrywania wiedzy przedstawionych na wykładzie lub innych z
dziedziny eksploracji danych.
Przykłady algorytmów do zaimplementowania w ramach projektu:
Algorytmy wyznaczania wzorców częstych;
Algorytmy wyznaczania zwięzłych reprezentacji wzorców częstych;
Algorytmy odkrywania wzorców sekwencyjnych;
Algorytmy grupowania danych;
Algorytmy odkrywania reduktów/reguł teorii zbiorów przybliżonych;
Algorytmy odkrywania zależności funkcyjnych i zależności
przybliżonych;
Algorytmy budowy i użycia klasyfikatorów;
Nie stawia się żadnych wstępnych warunków na narzędzia programistyczne,
które zostaną zastosowane do realizacji prac projektowych.

**Metody oceny:**

Warunkiem zaliczenia przedmiotu MED jest uzyskanie przez studenta pozytywnej oceny z każdego z trzech komponentów tego przedmiotu: części wykładowej (ocenianej na podstawie egzaminu), części projektowej (ocenianej na podstawie zrealizowanego oprogramowania i przeprowadzonych testów z jego użyciem, sprawozdania i prezentacji z realizacji projektu) oraz części laboratoryjnej (uznanej za zaliczoną, jeżeli wszystkie sprawdzające zadania laboratoryjne w liczbie pięciu są wykonane prawidłowo). Pozytywna ocena końcowa jest wyznaczana na podstawie średniej z oceny z egzaminu i oceny z projektu. Jeśli ocena z egzaminu jest niższa od oceny z projektu, wtedy ocena końcowa jest wyznaczana jako zaokrąglenie w dół tejże średniej. W przeciwnym przypadku jest wyznaczana jako zaokrąglenie w górę tejże średniej.

**Egzamin:**

tak

**Literatura:**

Han J., Kamber M., Pei, J., Data Mining: Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems, 3rd edition, Morgan Kaufmann, 2011
Morzy T., Eksploracja danych, Metody i algorytmy, Wydawnictwo Naukowe PWN, 2013
Fayyad U.M. , Piatetsky-Shapiro G. , Smyth P., ()Uthurusamy R. (eds.), Advances in Knowledge Discovery and Data Mining, AAAI, Menlo Park, California, 1996
Kryszkiewicz M., Concise Representations of Frequent Patterns and Association Rules, Prace Naukowe, Elektronika, Oficyna Wydawnicza Politechniki Warszawskiej, z. 142 (2002)
Communications of the ACM, November 1996, Vol. 39. No 11., 1996
Ganter B., Wille R., Formal Concept Analysis, Mathematical Foundations, Springer-Verlag, 1999
oraz szereg dostępnych na stronach Internetowych najnowszych publikacji z dziedziny eksploracji danych, o których wykładowca będzie informował na bieżąco w trakcie zajęć.

**Witryna www przedmiotu:**

https://usosweb.usos.pw.edu.pl/kontroler.php?\_action=katalog2/przedmioty/pokazPrzedmiot&prz\_kod=103B-INxxx-MSP-MED

**Uwagi:**

## Charakterystyki przedmiotowe

### Profil ogólnoakademicki - wiedza

**Charakterystyka MED\_W01:**

ma wiedzę dotyczącą odkrywania wzorców i zależności za pomocą metod eksploracji danych

Weryfikacja:

egzamin

**Powiązane charakterystyki kierunkowe:** K\_W06, K\_W08, K\_W09

**Powiązane charakterystyki obszarowe:** III.P7S\_WG.o, I.P7S\_WG

**Charakterystyka MED\_W02:**

ma podbudowaną teoretycznie wiedzę szczegółową dotycząca metod reprezentacji częstych wzorców i wnioskowania z ich użyciem

Weryfikacja:

egzamin

**Powiązane charakterystyki kierunkowe:** K\_W06

**Powiązane charakterystyki obszarowe:** I.P7S\_WG, III.P7S\_WG.o

**Charakterystyka MED\_W03:**

ma wiedzę z zakresu nowoczesnych technologii eksploracji danych

Weryfikacja:

sprawdzające ćwiczenia laboratoryjne

**Powiązane charakterystyki kierunkowe:** K\_W11

**Powiązane charakterystyki obszarowe:** I.P7S\_WG

### Profil ogólnoakademicki - umiejętności

**Charakterystyka MED\_U01:**

potrafi zaplanować proces odkrywania wiedzy, zrealizować go za pomocą zaimplementowanych przez siebie algorytmów oraz dokonać interpretacji uzyskanych wyników

Weryfikacja:

projekt

**Powiązane charakterystyki kierunkowe:** K\_U01, K\_U06, K\_U09, K\_U13

**Powiązane charakterystyki obszarowe:** I.P7S\_UK, I.P7S\_UW, III.P7S\_UW.1.o, III.P7S\_UW.3.o

**Charakterystyka MED\_U02:**

potrafi przedstawić plan, realizację i wyniki procesu odkrywania wiedzy oraz interpretację uzyskanych wyników w formie ustnej i pisemnej

Weryfikacja:

projekt

**Powiązane charakterystyki kierunkowe:** K\_U03, K\_U06

**Powiązane charakterystyki obszarowe:** I.P7S\_UK, I.P7S\_UW

**Charakterystyka MED\_U03:**

potrafi odkrywać wiedzę posługując się nowoczesnymi technologiami eksploracji danych

Weryfikacja:

sprawdzające ćwiczenia laboratoryjne

**Powiązane charakterystyki kierunkowe:** K\_U13

**Powiązane charakterystyki obszarowe:** I.P7S\_UW, III.P7S\_UW.3.o