“ ИЗВЛИЧАНЕ НА ИНФОРМАЦИЯ В ИНТЕРНЕТ”

Лектор: доц. д-р инж. Х.Вълчанов
Катедра: *КОМПЮТЪРНИ НАУКИ И ТЕХНОЛОГИИ*

Анотация:

Дисциплината има за цел да запознае студентите с принципите на извличане на информация в Интернет. Разглеждат се базовите понятия и методи на извличането на данни от документи, като се акцентира върху съвременните подходи и алгоритми на търсене на информация във Web пространството. Обърнато е внимание на въпросите, свързани с индексирането на информацията, модели за извличане на информация, рейтингуването и запитванията. Разглеждат се принципите на изграждане на търсещите машини във Web, както и на особеностите на съвременните системи за извличане на информация в Интернет.

Основни раздели на съдържанието:

1. Базови принципи на извличане на информацията (ИИ).
2. Архитектура на търсеща машина. Основни компоненти. Функциониране.
3. Извличане на web страници. Web Crawling. RSS feeds. Съхраняване на извлечените документи.
4. Обработване на текст. Оценка на резултантното множество.
5. Парсване на документ. Анализ на връзки.
6. Рейтингове и индексиране. Изграждане на индекси. Инвертни индекси. Компресиране.
7. Запитвания (queries). Трансфомации и прецизиране на запитванията. Извеждане на резултати.
8. Модели на извличане на информация.
9. Оценка на търсещите машини.
10. Класификация и клъстеринг. Разпознаване на спам.
11. Социално търсене. Тагове. Филтриране на документи.
12. Извличане на XML документи. Особености.
13. Системи за извличане на информация. LEXIS/NEXIS, SMART, Dialog, Dow Jones News/Retrieval, INQUERY.
14. Архитектура на търсещата машина на Google.
15. Извличане на мултимедийна информация.

Форма на изнасяне на учебното съдържание:

Лекции- включват общо 15 теми.

Лабораторни упражнения – провеждат се в специализирана компютърна зала, като се дава възможност на студентите да приложат на практика получените знания. Задават се задачи за самостоятелна подготовка. Провеждат се контролни работи по учебния материал.