



evropský  
sociální  
fond v ČR



EVROPSKÁ UNIE



MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání  
pro konkurenceschopnost

## INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Název projektu: Mezinárodní centrum pro informaci a neurčitost

Registrační číslo: CZ.1.07/2.3.00/20.0060

## Zpráva z účasti na konferenci

název konference: ICDM 2013  
datum konání: 7.12. - 10.12. 2013  
místo: Dallas, Texas, USA  
účastník konference: Mgr. Markéta Krmelová

### Stručný popis konference:

ICDM je přední mezinárodní konferencí zaměřenou na Data mining. Tato konference pokrývá široké spektrum témat v oblasti Data miningu včetně algoritmizace, vytváření software, systémů a aplikací.

### Témata konference:

- Data mining
- Machine learning
- Text mining
- Pattern recognition
- Data mining velkých dat
- Složitost metod Data miningu
- Dotazovací jazyky
- Předzpracování dat
- Redukce dat
- Soft computing (neuronové sítě, fuzzy logika, rough sets)
- Atd.

### Zajímavá čísla:

|                            |                                       |
|----------------------------|---------------------------------------|
| Počet zaslaných příspěvků: | 809                                   |
| Počet přijatých příspěvků: | 94 regular papers and 65 short papers |
| Počet paralelních sekcí:   | 4                                     |
| Acceptance rate:           | 19.65%                                |

## **Zajímavé přednášky:**

### **A. Plenární přednášky**

#### *Large-scale Learning in Computational Advertising*

*Jianchang (JC) Mao*

Přednáška se zabývala jedním z nejrychleji se rozvíjející odvětví reklamy, tzv. online advertising. Do tohoto odvětví spadá optimalizace vrácených příspěvků, které dostane uživatel od vyhledavače. Jelikož pro každý dotaz může vyhledavač vrátit tisíce relevantních výsledků, je nutné použít metody šité na míru velkým datasetům. Bylo zde představeno několik problémů souvisejících s velkým objemem dat a také modifikace již známého EM algoritmu pro minimalizaci KL-divergence.

### **B. Řádné přednášky**

#### *Classifying Spam Emails using Text and Readability Features*

*Rushdi Shams and Robert E. Mercer*

V tomto příspěvku byla představena definice SPAMu a metody pro jejich klasifikaci. Doposud známé metody emaily klasifikovaly rozbořením hlavičky nebo obsahu jednotlivých emailů. Spameři však tyto metody dokáží obejít. Nově představená metoda klasifikuje emaily na základě obsahu dříve klasifikovaných spam emailů.

#### *Noise-Resistant Bicluster Recognition*

*Huan Sun, Gengxin Miao and Xifeng Yan*

V tomto příspěvku byla představena nový model biclusteringu – AutoDecoder. V tomto modelu se spojují znalosti z machine learningu a neuronových sítí. Tato metoda byla představena na čtyřech reálných datasetech a několika syntetických. Nejlépe si tato metoda vedla na datech, které obsahovaly méně než 15% šumu.

#### *Discovering Non-Redundant Overlapping Biclusters on Gene Expression Data*

*Duy Tin Truong, Roberto Battiti, Mauro Brunato*

V tomto příspěvku byl představen nový algoritmus pro generování K překrývajících se biclusterů, jejichž překrytí je menší než stanovený mez. Tento algoritmus byl prezentován na reálných datasetech, na kterých bylo ukázáno, že metoda dává větší a kvalitnější biclustery, než doposud známé metody.

### **Prezentovaný příspěvek**

Na této konferenci byl prezentován článek napsaný společně s R. Bělohlávkem:

Beyond Boolean Matrix Decompositions: Toward Factor Analysis and Dimensionality Reduction of Ordinal Data.

Booleovské faktorizaci matic (BMF), nebo také dekompozici, je věnována značná pozornost v oblasti Data miningu a to jak pro její přímé využití při analýze dat, ale také proto, že hraje roli v porozumění Booleovských dat. V tomto článku tvrdíme, že výzkum by měl jít až za hranice Booleovského případu směrem k obecnějšímu typu dat, jako jsou ordinální data.

Technicky toto rozšíření představuje výměnu dvouprvkové Booleovy algebry za obecnější strukturu. Tato výměna přináší řadu netriviálních problémů. V první řadě formulujeme problém, zkoumáme existující literaturu a uvádíme názorný příklad. Zadruhé uvádíme nové teoretické výsledky související s dekompozicí ordinálních dat. Tyto výsledky napomáhají pochopení geometrie rozkladů a identifikují části vstupních matic, na které je vhodné se při výpočtu faktorů zaměřit. Zatřetí navrhujeme dva nové algoritmy založené na těchto výsledcích spolu s experimentálním hodnocením.

### **ICDM 2013 Community Meeting**

V rámci IFSA konference, proběhlo i setkání ICDM community meeting. Všichni účastníci konference se tohoto setkání mohli zúčastnit. Byla zde představena budoucnost ICDM konference, její vývoj během posledních let, včetně změn trendů témat příspěvků na této konferenci.

### **Shrnutí konference (perspektivní témata apod.)**

Na konferenci byly prezentovány příspěvky z oblasti Data miningu, zejména zaměřené na aplikace v oblasti webových technologiích, sociálních sítích, bioinformatice, medicíně a jiných odvětvích informačních technologií – soft computingu.