Przegląd konferencji ICML, SIGIR i ICDAR w kontekście bibliotek cyfrowych

Zebrał i podsumował: Tomasz Kuśmierczyk
ICM UW, MIM UW

Warszawa, 2011-09-15

Po co ?

- Co jest modne
- Co już ktoś zrobił
- Gdzie szukać rozwiązań
- Gdzie szukać pomysłów
- Mój sposób na wyrabianie godzin w pracy

Plan prezentacji

1. Krótkie omówienie źródeł
2. ICML
3. SIGIR
4. ICDAR

Sposób omawiania

- Podział wg istotności
- Wyjaśnienie
- Przykłady

Źródła

- ICML '2010
- SIGIR '2011
- ICDAR '2009
- Więcej na: https://wiki.yadda.icm.edu.pl/yadda/Content_Analysis_service/State_of_the_art

- http://en.wikipedia.org

ICML

- International Conference on Machine Learning
- Najstarsza (od 1984) i najważniejsza o Uczeniu Maszynowym
- Zagadnienia teoretyczne
- Dużo matematyki, modelowania itp.
- Ale też: liczne przykłady zastosowań

SIGIR

- ACM SIGIR Conf on Information Retrieval
- Od 1971
- Jedna z najważniejszych o IR
- Wszystkie aspekty IR
- Zagadnienia teoretyczne
- Zagadnienia praktyczne

SIGIR

- ...
- Modyfikacje istniejących podejść
- Zastosowanie narzędzia do nowego problemu
- Kilka pomysłowych rozwiązań

SIGIR a ICML

- ICML
  - Uczenie maszynowe
  - Teoria
- SIGIR
  - Dokumenty
  - Praktyka i teoretyczna podbudowa zastosowań

ICDAR

- Intl Conf on Document Analysis and Recognition
- Od 1991
- Podobno Ważna
- Typowe problemy
- Zastosowania istniejących narzędzi do nowych danych
- Czasem modyfikacja podejścia

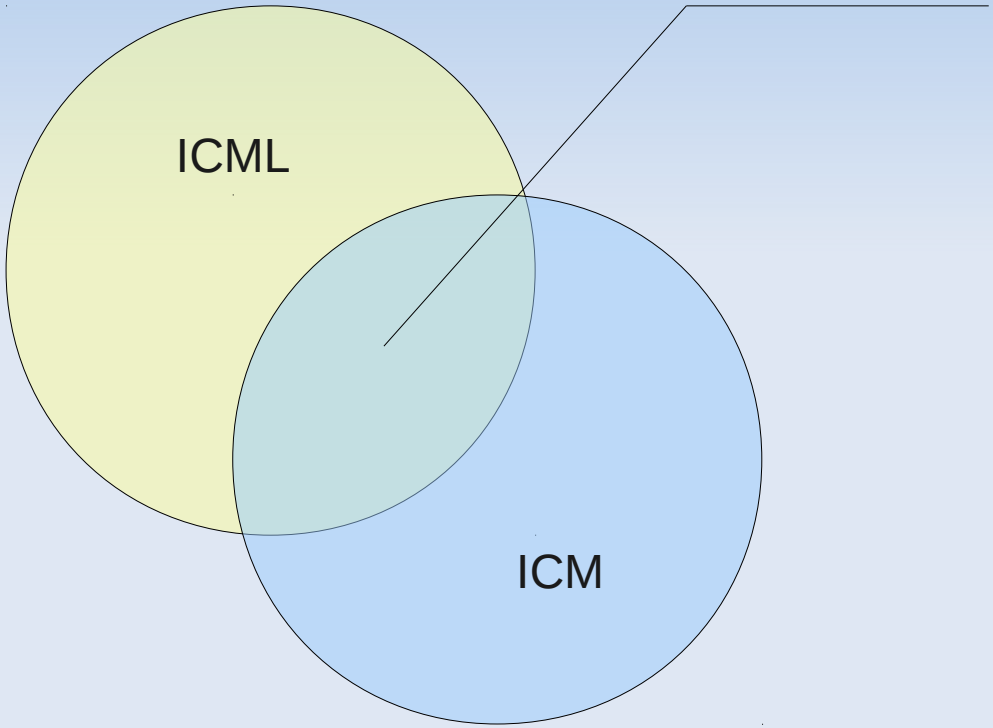
ICML


HAIFA
ICML
ISRAEL
2010
International
Conference on
Machine
Learning
Haifa, Israel
June 21 - 24

ICML: Duże rzeczy

Lp.	Temat	Liczba prac
1	Reinforcement Learning	16
2	Clustering + Graph Clustering	12
3	Features & Kernels Feature and Kernel Selection/Kernels/Exploration and Feature Construction	12

Duże rzeczy a praktyka ICM

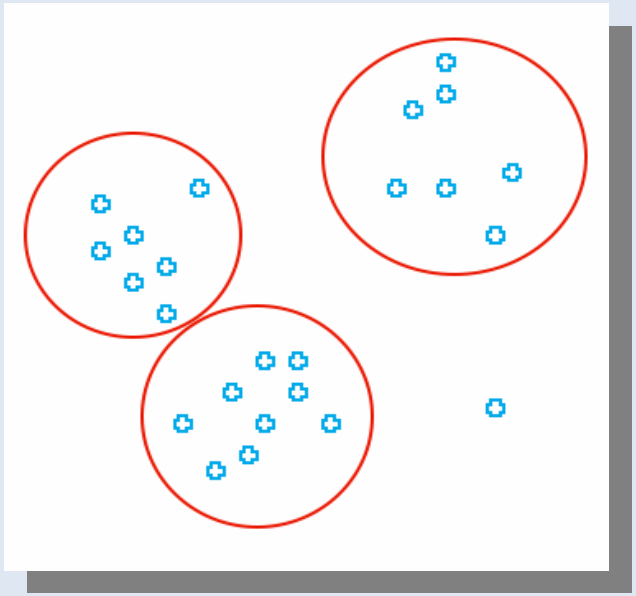


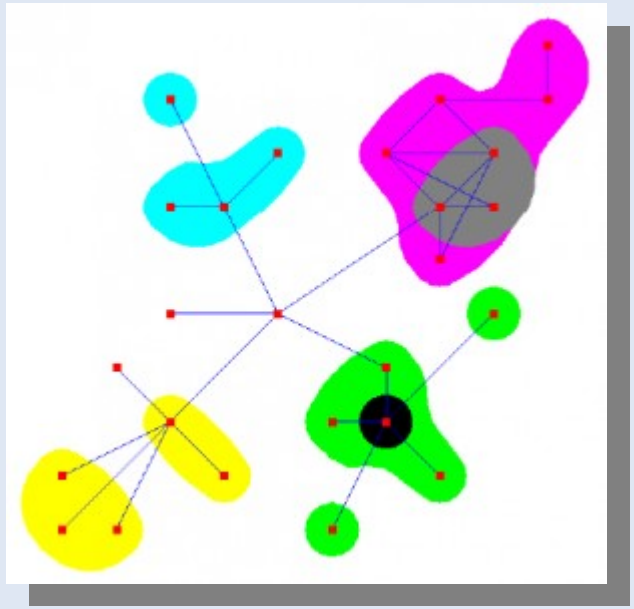
★★★ Clustering + Graph Clustering

★★★ Features & Kernels
Feature and Kernel Selection/Kernels/Exploration and Feature Construction

Clustering + Graph Clustering

- Podstawowe narzędzie ML
- Clustering vs. Graph Clustering





Clustering + Graph Clustering: co można robić

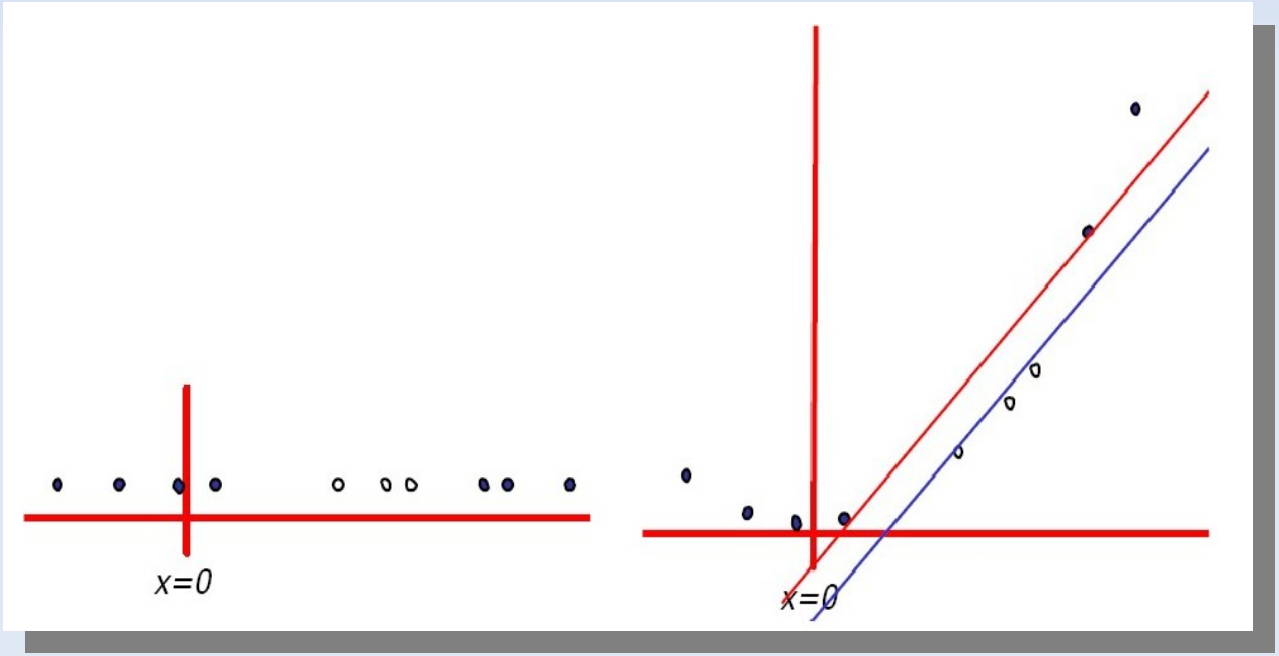
- Nowa heurystyka dla grafów
- Adaptacja pomysłu na potrzeby clusteringu
- Analiza i porównanie istniejących metod dla specyficznych danych

Clustering + Graph Clustering: przykładowa publikacja

- ***Finding Planted Partitions in Nearly Linear Time using Arrested Spectral Clustering***
  - na wejściu graf podobieństwa o pewnej charakterystyce
  - propozycja algorytmu w narzuconej złożoności
  - dowód teoretyczny, że znajduje klastry o zadanej charakterystyce (np. ”wystarczająco duże”)
  - porównanie z innymi algorytmami

Features & kernels: czym są kernele I

- Transformacja danych do wysokowymiarowych przestrzeni



Źródło: **http://www.mimuw.edu.pl/~son/ML2011/**

Features & kernels: czym są kernele II

- Często zamiast transformacji jest funkcja licząca iloczyn skalarny w nowej przestrzeni
- *Kernel methods* są szeroko stosowane:
  - SVM
  - LDA
  - PCA
  - ...

Features & kernels: co można robić

- Wybór i uczenie się cech
- Wybór i uczenie się kerneli
- Użycie *kernel methods* w nowym zagadnieniu
- …

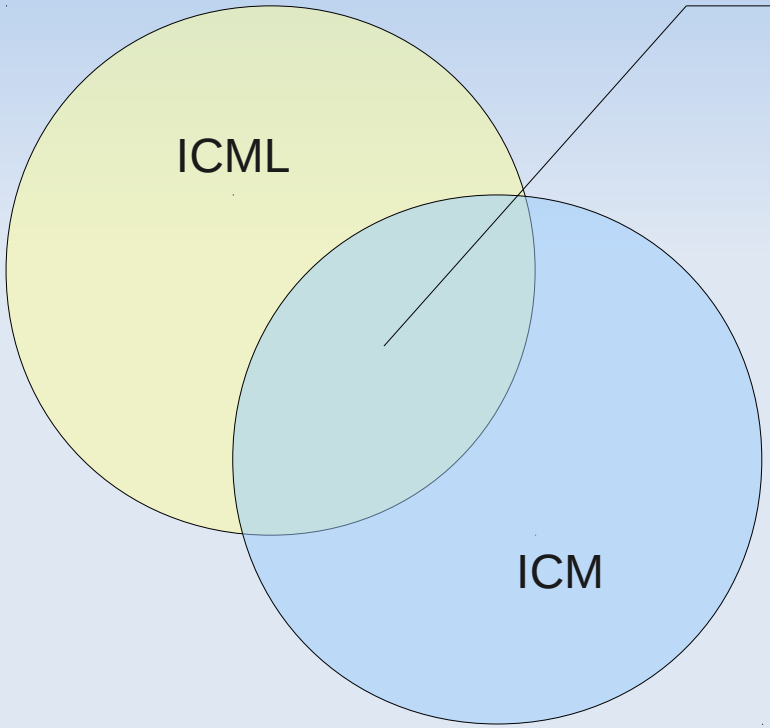
Features & kernels: przykładowa publikacja

- ***Online Streaming Feature Selection***:
  - problem:
    - nie wiemy ile dane mają atrybutów
    - atrybuty przychodzą w strumieniu
    - w danym momencie chcemy mieć wybrany najlepszy podzbiór możliwy podzbiór cech
  - rozwiązanie: nowy algorytm
  - porównanie z istniejącymi metodami

ICML: Ważne kwestie

Lp.	Temat	Liczba prac
4	Deep Learning	8
5	Dimensionality Reduction	8
6	Semi-Supervised Learning	8
7	Online Learning/Active Learning	8
8	Graphical Models and Bayesian Methods	8
9	Topic Models and Matrix Factorization / Latent-Variable Models	8

Ważne kwestie a praktyka ICM



★★★ Topic Models / Latent-Variable Models

☆★★ Graphical Models and Bayesian Methods

☆☆★ Dimensionality Reduction

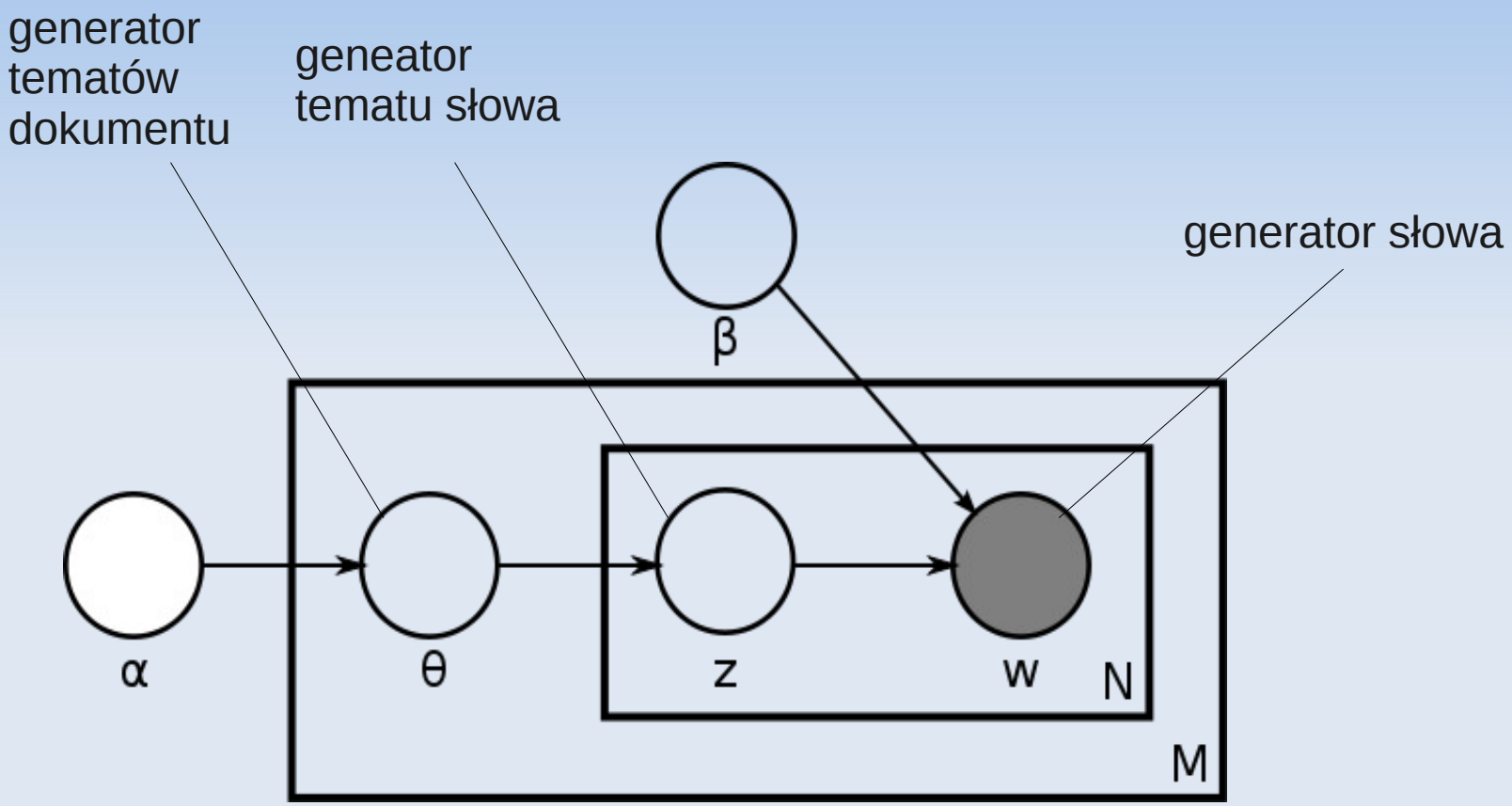
Latent-Variable Models

- model statystyczny
- ukryte zmienne
- obserwowalne wskaźniki
- dział interesujący dla ICM:
  - Latent semantic analysis (~Topic models) – dokument jako zbiór konceptów (znaczeń)

Topic Models na przykładzie Latent Dirchlet Allocation I

- Założenie: korpus dokumentów jest generowany w procesie opartym o kilka zmiennych losowych:
  - generowany jest pewien rozkład tematów dla każdego z dokumentów
  - dla każdego ze słów dokumencie generowany jest temat
  - z rozkładu słów nad dokumentem generowane jest pojedyncze słowo w dokumencie

Topic Models na przykładzie Latent Dirchlet Allocation II

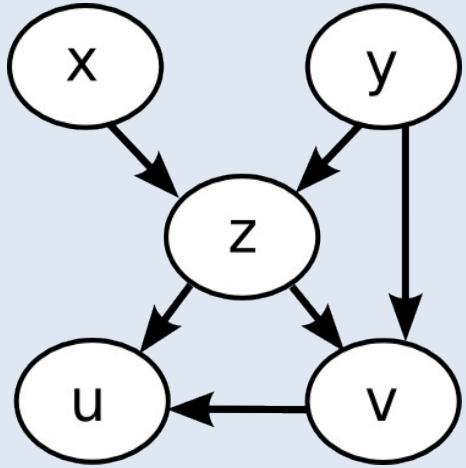


Latent Semantic Analysis/Models przykładowe publikacje

- ***Spherical Topic Models*** – ulepszenie LDA
- ***A Language-based Approach to Measuring Scholarly Impact*** – identyfikacja wpływowych dokumentów poprzez analizę tematów (w znaczeniu: topics) w czasie
- ***ILDA: Interdependent LDA Model for Learning Latent Aspects and their Ratings from Online Product Reviews*** – ekstrakcja ocen produktów z opinii na forach

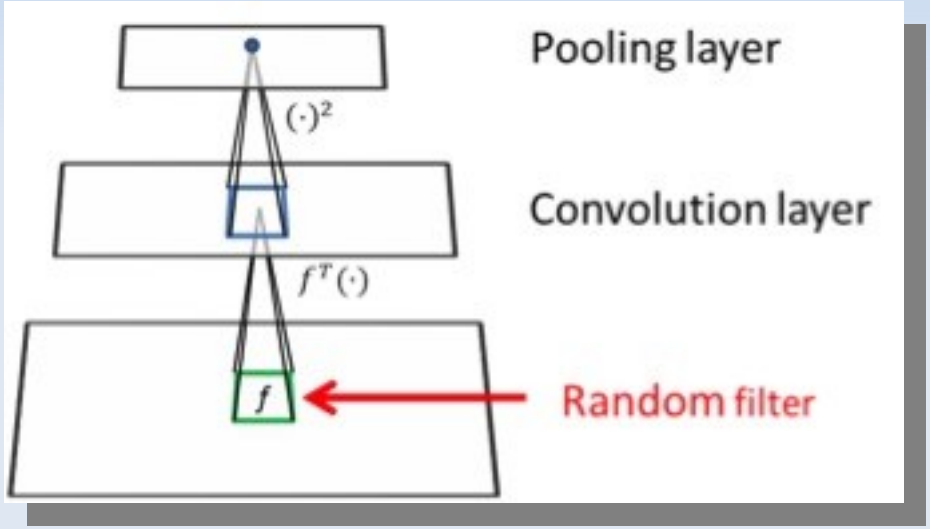
Graphical Models and Bayesian Methods

- Model probabilistyczny
- Reprezentowany jako graf
- Koduje warunkową niezależność
- Przykłady:
  - Sieci Bayesowskie
  - Sieci Markowa



Deep Learning I

- Wiele poziomów nieliniowej transformacji cech np. piksele, krawędzie, kształty
- Proces uczenia uwzględnia warstwowość
- Subprocedury zamiast budowania ”grubych” klasyfikatorów



Deep Learning II

- Przykładowa publikacja:
  - ***3D Convolutional Neural Networks for Human Action Recognition***
- Więcej na: https://wiki.yadda.icm.edu.pl/yadda/Content_Analysis_service/Hot_science_topics/ML

Dimensionality Reduction

- W najprostszym przypadku:
  - Na wejściu dane wielowymiarowe
  - Na wyjściu małowymiarowe
  - Minimalna strata informacji

- Publikacje czysto teoretyczne

ICML: Drobnica I

Lp.	Temat	Liczba prac
9	Ensemble Methods	4
10	Statistical Relational Learning	4
11	Large-Scale Learning and Optimization	4
12	Matrix Factorization and Regularization	4
14	Risk estimation and Cost-sensitive Learning	4
15	Causal Inference	4

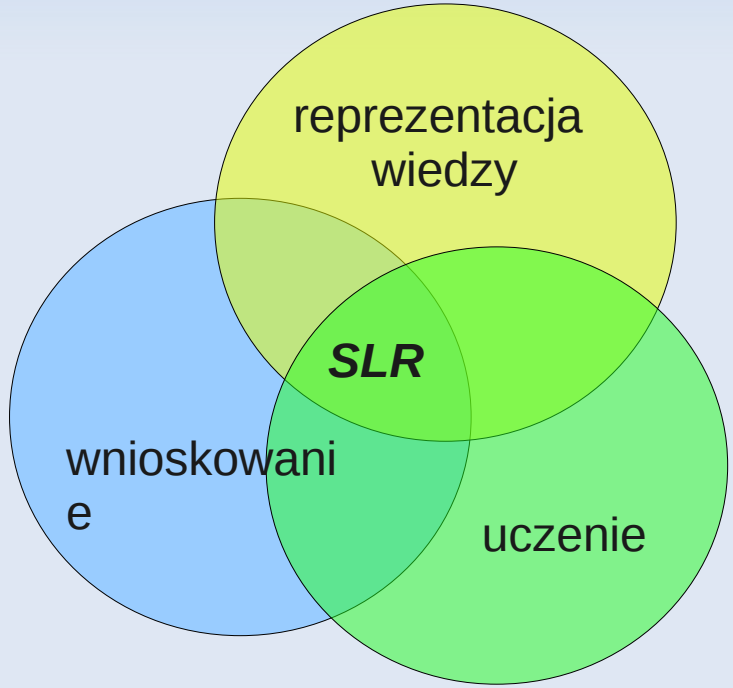
Ensemble methods

- Mamy wiele klasyfikatorów
- Chcemy mieć jeden
- Jeden ale lepszy od każdego

- Publikacje: nowe metody łączenia, nowe zastosowania
- Przykład: ***Boosting for Regression Transfer***

Statistical Relational Learning I

- Własności rozważanych struktur:
  - niepewność (ang. uncertainty) → model statystyczny
  - relacyjność
  - złożność



Statistical Relational Learning II

- Relacje w FOL (zdania + kwantyfikacje)
- Całość zamodelowana probabilistycznie
  - Sieci Bayesowskie
  - Sieci Markowa

Statistical Relational Learning przykład

- ***Active Learning for Networked Data***
  - klasyfikacja danych w sieci
  - standardowe metody
  - nowość: active learning

ICML: Drobnica II

Lp.	Temat	Liczba prac
16	Large Margin Methods	4
17	Compact Representations	4
18	Gaussian Processes	4
19	Multi-Agent Learning	4
20	Time-Series Analysis	4
21	Multi-Label and Multi-Instance Learning	4
22	Learning from humans	4

Compact Representations

- *Learning Fast Approximations of Sparse Coding*
- *Submodular Dictionary Selection for Sparse Representation*
- *Proximal Methods for Sparse Hierarchical Dictionary Learning*
- *Sequential Projection Learning for Hashing with Compact Codes*

Multi-Instance Learning

- Grupy przykładów
- Grupy poetykietowane +/-
- ”-” = wszystkie przykłady w grupie ”-”
- ”+” = przynajmniej jeden przykład w grupie ”+”
- Cel: nauczyć się +/- dla pojedynczych przykładów

Multi-Label Learning

- Dane z wieloma etykietami
- Zastosowanie np. kategoryzacja tekstu
- Przykład: ***Graded Multilabel Classification: The Ordinal Case***
  - obiekty mają ważoną przynależność do klas
  - dwa nowe algorytmy jak przejść do tradycyjnej klasyfikacji
  - modyfikacja tradycyjnych metryk z multi-label learningu aby uwzględniały ważenie

ICML: Pył

Lp.	Temat	Liczba prac
25	Multi-Task and Transfer Learning	3
26	Ranking and Preference Learning	3
27	Structured Output Learning	3
28	Learning Theory	3

Dziobak

- Ssak który składa jaja
- Mieszka w Australii
- Pozwala na zainteresowanie widowni
- Odrywa od tematu

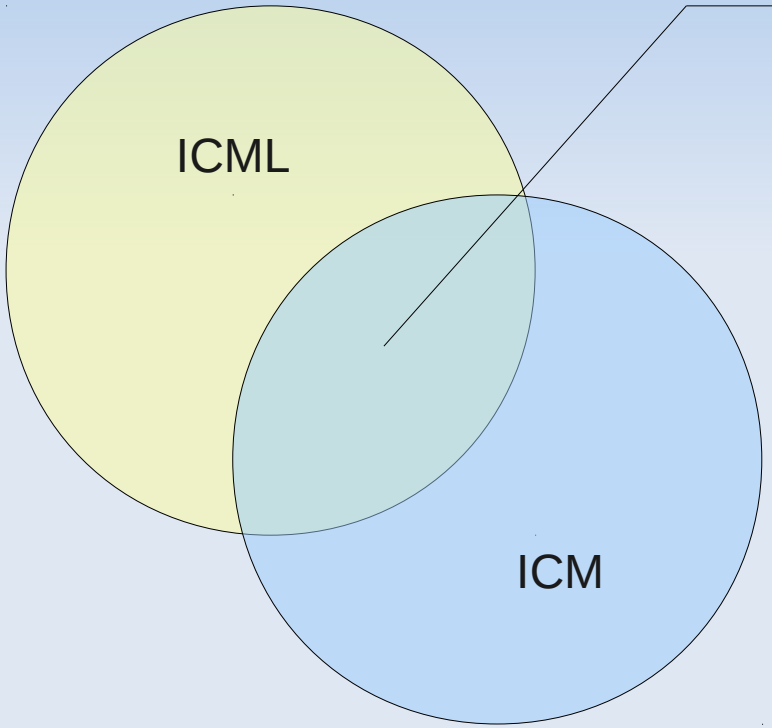
SIGIR


SIGIR
Beijing
2011
ACM
SIGIR 2011
July 24-28, 2011
The 34th Annual International ACM SIGIR Conference

SIGIR: duże rzeczy

Lp.	Temat	Liczba prac
1	Queries (Query Analysis / Web Queries / Query suggestions)	16
2	Collaborative filtering / Recommender systems	12
3	Users	8
4	Retrieval models	6

Duże rzeczy w SIGIR a ICM



Queries

Recommender systems

Users

Retrieval models

Queries I

- Polepszanie wyników zapytań np. poprzez wprowadzenie dywersyfikacji wyników
  - ***Incremental Diversification for Very Large Sets: a Streaming-based Approach***
    - Dywersyfikacja wyników poprawia szanse na trafienie interesującego dokumentu
    - Zapronowano nową heurystykę radzącą sobie z danymi strumieniowymi (bez dostępu do całego korpusu)

Queries II

- Polepszanie wyników zapytań np. poprzez zastosowanie wielu modeli
  - ***Intent-Aware Search Result Diversification***
    - Analizujemy zapytanie pod różnymi aspektami
    - Dla każdego aspektu inny model
    - Inny model daje inne wyniki
    - Uzyskuje się dywersyfikację

Queries III

- Wyszukiwanie w nowych dziedzinach np.:
  - ***People Searching for People: Analysis of a People Search Engine Log*** – wyszukiwanie ludzi
- Uwzględnienie nowych aspektów danych np.:
  - ***Social Annotation in Query Expansion: a Machine Learning Approach*** – wyniki z sieci społecznościowych

Queries IV

- Podpowiadanie zapytań
  - ***Automatic Boolean Query Suggestion for Professional Search***
    - Środowiska specjalistyczne (np. Poszukiwanie patentów) – charakterystyczny sposób wyszukiwania:
      - Wiele zapytań dotyczących tematu
      - Sprawdzenie wielu wyników
    - Generowane są zapytania boolowskie w oparciu o drzewa poetykietowanych dokumentów

Systemy rekomendujące I

- ICML: Ranking and Preference Learning
- SIGIR: Collaborative filtering / Recommender systems

Systemy rekomendujące II

- Uczenie się preferencji użytkownika
- Przewidywanie co go zainteresuje
- Włączanie nowych aspektów w systemy rekomendacyjne (np. geografia, czas)
- Propozycje nowych modeli / metod

Systemy rekomendujące: przykładowa publikacja I

- ***Collaborative Competitive Filtering: Learning Recommender Using Context of User Choices***
  - Rozważany jest biznes on-line
  - Zaproponowano model który uwzględnia kontekst akcji/decyzji użytkownika
  - Pokazano zastosowanie modelu do uczenia się preferencji użytkownika

Systemy rekomendujące: przykładowa publikacja II

- ***Learning Relevance from Heterogeneous Social Network and Its Application in Online Targeting***
  - propozycja modelu łączenia informacji o preferencjach z różnych źródeł
  - źródła mają różną ale nieznaną ważność
  - test na reklamach Facebook-a

Systemy rekomendujące: przykładowa publikacja III

- ***Label Ranking Methods based on the Plackett-Luce Model***
  - dany jest model probabilistyczny rankingowanych danych
  - zaproponowano dwie metody budowania rankingu

Users I

- analiza zachowań ludzi w związku z wyszukiwaniem / interakcją z systemami
- bardziej modelowanie zachowań niż wymyślanie
- taka socjologia / kogniwistyka
- możliwe wszystko co jest związane z użytkownikami systemów

Users: przykładowa praca I

- ***Understanding Re-finding Behaviour in Naturalistic Email Interaction Logs***
  - analizowane są logi z klientów pocztowych
  - sformalizowano i zamodelowano zachowań użytkowników w takim środowisku
  - zaproponowano metodę wyszukiwania i analizy akcji typu: ponowne wyszukanie i użycie poczty

Users: przykładowa praca II

- ***Measuring Improvement in User Search Performance Resulting From Optimal Search Tips***
  - Poprawa wyników przez dodatkowe wskazów w trakcie wyszukiwania

Retrieval models

- Modele dokumentów
- Umożliwają efektywne pozyskiwanie informacji
- Efektywne względem zapytań

Retrieval models historyczne podejścia

- Tagowanie tematyczne (+ informacja o semantyce) dokumentów
- Porównywanie wektorów cech dokumentów
- Podejścia probabilistyczne
- Modele języka
- PageRank

Retrieval models
jakie prace?

- Nowe podejścia / pomysły jak poprawić wyniki
- Poprawienie istniejących modeli np. po pokazaniu gdzie wcześniejsze nie wystarczają
- Modyfikacje istniejących modeli o nowe aspekty np. czas, geografia

Retrieval models
przykładowa publikacja I

- ***Enhancing Ad-hoc Relevance Weighting Using Probability Density Estimation***
  - dwa poglądy na istotność długości dokumentu
  - wprowadzono probabilistyczny model długości dokumentów
  - w oparciu o ten model zmodyfikował klasyczne podejście BM25

Retrieval models przykładowa publikacja II

- ***Estimation Methods for Ranking Recent Information***
  - Modyfikacja *Query Likelihood Model*
  - Dodanie informacji o aktualności dokumentu
- *Query likelihood model:*
  - model języka budowany dla każdego dokumentu
  - daje prawdopodobieństwo dopasowania dokumentu do zapytania (*query*)

SIGIR: Drobnica I

Lp.	Temat	Liczba prac
5	Web IR	4
6	Vertical & Entity Search	4
7	Test collections	4
8	Linguistic Analysis	4
9	Learning to Rank	4
10	Indexing	4
11	Efficiency	4
12	Content Analysis	4
13	Communities	4

SIGIR: Drobnica II

Lp.	Temat	Liczba prac
14	Summarization	3
15	Social Media	3
16	Personalization	3
17	Multimedia IR	3
18	Multilingual IR	3
19	Latent Semantic Analysis	3
20	Image Search	3
21	Effectiveness	3
22	Clustering	3
23	Classification	3

Linguistic Analysis

- Analiza językowa dokumentów:
  - Nowe algorytmy przetwarzania np. stemmingu
  - Rozpoznawanie znaczenia słów np. toponimy
  - Wzbogacanie reprezentacji o np. wyniki tłumaczenia
- ***Improved Video Categorization from Text Metadata and User Comments***

Content Analysis

- Segmentacja dokumentów
- Analiza ważności / znaczenia sekcji
- ***Detecting Outlier Sections in US Congressional Legislation***
  - formalne dokumenty są długie i napisane trudnym językiem
  - próbujemy podzielić je na sekcje i wyróżnić najważniejsze

Communities

- Wyszukiwanie i analiza społeczności
- ***Mining Topics on Participations for Community Discovery***
  - Zaprojektowano powiązania między autorami
  - Zbudowano graf
  - Zaadaptowano nową metodę do wyszukiwania podgrafów

Summarization

- Budowanie podsumowań i zestawień z dokumentów i korpusów dokumentów
  - **Summarizing the Differences in Multilingual News**
    - porównanie news-ów angielskojęzycznych i chińskojęzycznych
    - podsumowanie najważniejszych różnic

Multilingual IR

- Dokumenty w wielu językach
  - poszukiwanie różnic
  - ujednoznacznianie
  - Wzbogacanie
- ***No Free Lunch: Brute Force vs. Locality-Sensitive Hashing for Cross-lingual Pairwise Similarity***

Learning to Rank

- *A Cascade Ranking Model for Efficient Ranked Retrieval*
  - Zbudowanie dobrego rankingu wymaga przejrzenie wielu obiektów i z wieloma cechami
  - Jest to jednak bardzo wolne
  - Podejście hierarchiczne:
    - Szybsze modele robią preselekcję w kilku fazach
    - Wolniejsze ale lepsze następnie budują ranking

Image Search

- Wyszukiwanie obrazów i w obrazach
- ***Integrating Hierarchical Feature Selection and Classifier Training for Multi-Label Image Annotation***
  - przegląd cech obrazków - propozycja algorytmu hierarchicznego wyboru cech
  - cechy są następnie agregowane
  - budowany jest wielo-etykietowy klasyfikator

ICDAR


10th International Conference on Document Analysis and Recognition
icdar2009
July 26-29, 2009
Universitat Autònoma BARCELONA
CATALONIA-SPAIN
IAPR
Centre de Visió per Computador
UAB
Universitat Autònoma de Barcelona

ICDAR I

Lp.	Tematyka
1	Character/Handwritting analysis & recognition
2	Seal/icons recognition
3	Text detection/identification (from scans, images, videos etc.)
4	On/Off-Line Signature Verification
5	Writer identifcation based on Handwriting
6	Document zones analysis
7	Text features extraction
8	Documents classification

ICDAR II

- Page Segmentation and Layout Analysis
- Scientific/Historical Document Recognition
- Ground-truthing
- Text Line Segmentation

Seal/icons recognition

- ***Seal Detection and Recognition: An Approach for Document Indexing***
  - Zmodyfikowana transformacja Hough-a do detekcji regionów
  - Cechy odporne na obroty i skalowania
  - SVM do klasyfikacji znaków
  - Odległości i kąty do detekcji wzajemnej orientacji
  - Na podstawie tekstu wyliczany hash i klasyfikacja

Document zones analysis

- ***Hybrid Page Layout Analysis via Tab-Stop Detection***
  - nowy hybrydowy algorytm
  - detekcja tab-stop do rozpoznania kolumn (bottom-up)
  - Top-down do detekcji kolejności słów

Text detection/identification

- *Real-Time Camera-Based Recognition of Characters and Pictograms*
  - W oparciu o standardowe metody
  - Wzbogacone o algorytm głosowania i haszowanie
  - Może działać w czasie rzeczywistym

Text features extraction

- ***A New Block Partitioned Text Feature for Text Verification***
  - Cel: weryfikacja czy region faktycznie jest tekstem
  - Zaproponowano dwie nowe cechy
  - SVM do klasyfikacji

Documents classification

- ***Graph b-Coloring for Automatic Recognition of Documents***
  - nowe podejście do zarówno rozponawania układu dokumentu jaki i klasyfikacji
  - rozwiązanie oparte o kolorowanie grafu
  - ...

Dziękuję za uwagę. Pytania?