

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:	ACM SIGIR 2011 (The 34th Annual Int. ACM SIGIR Conference on Research and Development in Informational Retrieval)
datum konání:	24.7. - 28. 7. 2011
místo:	Peking, Čína
účastník konference:	Mgr. Jan Outrata, Ph.D.

Stručný popis konference:
ACM SIGIR je hlavní konferencí, která se zabývá reprezentací, získáváním, vyhledáváním a analýzou informací z rozsáhlých elektronických zdrojů (zejména internetu a WWW). Pravidelně se zde setkávají hlavní a nejcitovanější osobnosti z této oblasti informatiky. Konference je vedle sdružení ACM sponzorována největšími firmami (Microsoft, Google, Yahoo, IBM, HP) a významnými organizacemi (Baidu, eBay, Yandex) v oboru. Mezi účastníky jsou jak vědci z akademické sféry, tak vědci z výzkumných oddělení vůdčích firem (Microsoft Research, IBM Research).

Zajímavá čísla

Acceptance rate:	19,9%
Počet účastníků:	přes 800
Počet prezentovaných příspěvků:	231 (108 článků, 89 posterů, 15 demonstrací, 8 workshopů a 11 tutoriálů)
Procenta počtu článků dle zemí:	48 % USA, 17 % Čína, 6 % Německo, 29 % v součtu ostatní
Počet plenárních přednášek:	2
Počet paralelních sekcí	3

Zajímavé přednášky

A. Plenární přednášky

Materiály k přednáškám jsou dostupné na http://www.sigir2011.org/keynotes.htm**.**

Qi Lu (Microsoft): Future of the Web and Search.

Přednášející přestavil svoje myšlenky a vize o tom, kam směřuje současný Web a jakým směrem se budou vyvíjet vyhledávací technologie pro Web nový. Tento byl představen jako reorganizovaný pro vyhledávání založeném na znalosti identity, sociálních interakcí, lokace a předchozího chování uživatelů (nejen na webu) a porozumění tak tomu, co uživatelé chtějí. V tomto směru byla prezentována některá specifika Microsoftu, zahrnující např. využití metod information retrieval, a výzvy a možnosti pro akademický výzkum v této oblasti.

ChengXiang Zhai (University of Illinois at Urbana-Champaign): Beyond Search: Statistical Topic Models for Text Analysis.
Přednáška pojednávalo o využití statistických modelů pro analyzování velkého množství informací získaných z vyhledávání a nalezení tak relevantní informace. Modely zahrnovaly metody analýzy a extrakce témat a jejich vzorů v textu (konkrétně např. Různé varianty probabilistic latent semantic analysis) a představeny byly i ukázkové aplikace modelů v různých manažerských oblastech (comparative summarization, contextual topic trend analysis, event impact analysis).

B. Řádné přednášky

Abstrakty přednášek jsou dostupné na http://www.sigir2011.org/papers.asp**.**

Y. Ganjisaffar, R. Caruana, C. V. Lopes: Bagging Gradient-Boosted Trees for High Precision, Low Variance Ranking Models.
V přednášce bylo představeno vylepšení rankovacího modelu LambdaMART pomocí kombinace metod pro redukci odchylky (tzv. boosting) a rozptylu (tzv. bagging).

M. Ji, J. Yan, S. Gu, J. Han, X. He, W. V. Zhang, Z. Chen: Learning Search Tasks in Queries and Web Pages via Graph Regularization.
Přednáška pojednávalo o klasifikaci výsledků dotazů i dotazů samotných (s využitím logů dotazů) pomocí konstrukce grafu mezi výsledky a dotazy, a o novém algoritmu regularizace grafu.

L. Wang, J. Lin, D. Metzler: A Cascade Ranking Model for Efficient Ranked Retrieval.
V přednášce byl představen rankovací model pomocí kaskády rankovacích funkcí, který vede ke zlepšení jak rozhodovací účinnosti top-k dotazů tak rychlosti pro vyhledávání, a algoritmus pro naučení tohoto modelu.

A. Tian, M. Lease: Active Learning to Maximize Accuracy vs. Effort in Interactive Information Retrieval.
Přednáška pojednávala o interaktivním information retrieval systému s využitím strategie aktivního učení, kdy (simulované) hodnocení uživatelem nejvíce nejistě automaticky ohodnocených dokumentů funguje jako zpětná vazba pro zlepšení hodnocení.

D. Fernandes, E. S. de Moura, A. S. da Silva, B. Ribeiro-Neto, E. Braga: A Site Oriented Method For Segmenting Web Pages.
V přednášce byla představena plně automatická metoda pro segmentaci webových stránek a ukázáno její využití ve vyhledávačích, kde se výsledky vyhledávání blíží výsledkům při ruční segmentaci.

H. Ma, Ch. Liu, I. King, M. R. Lyu: Probabilistic Factor Models for Web Site Recommendation.

V přednášce byl představen pravděpodobnostní model pro doporučování personalizovaného webového obsahu založený na technikách redukce dimenze pomocí faktorizace reálných matic, který by měl být lineárně škálovatelný.

K. Zhou, S.-H. Yang, H. Zha: Functional Matrix Factorizations for Cold-Start Recommendation.
V přednášce byla představena metoda pro profilaci nových uživatelů (tzv. cold-start recommendation) pomocí prvotního dotazování uživatele, která kombinuje konstrukci rozhodovacího stromu na základě odpovědí uživatele a využívání dat z existujících profilů.

Y. Li, B.-J. P. Hsu, Ch. Zhai, K. Wang: Unsupervised Query Segmentation Using Clickthrough for Information Retrieval.
V přednášce byl představen pravděpodobnostní model pro segmentaci dotazu pro vyhledávání využívající historie odkazů (tzv. clickthrough data) a diskutována interpretace výsledků segmentace.

M. Ageev, Q. Guo, D. Lagun, E. Agichtein: Find It If You Can: A Game for Modeling Different Types of Web Search Success Using Interaction Data.
Přednáška pojednávala o modelování chování úspěšně vyhledávajícího uživatele využívajícím herní strategii pro vyhodnocení úspěšnosti a poskytující efektivní predikci úspěšnosti. Článek byl oceněn jako nejlepší (best paper award).

N. Moraveji, D. Russell, J. Bien, D. Mease: Measuring Improvement in User Search Performance Resulting From Optimal Search Tips.
Přednáška pojednávala o experimentální studii kvantitativního měření zlepšení uživatelů ve vyhledávání obsahu spočívající v řešení úkolů nejdříve se zobrazením tipů (a naučení se tak) a za týden bez tipů. Studie má význam pro designéry vyhledávačů.

Y. Lin, H. Lin, S. Jin, Z. Ye: Social Annotation in Query Expansion: A Machine Learning Approach.
Přednáška pojednávala o využití sociálních anotací vyhledávaných dokumentů pro rozšíření dotazu pro vyhledávání pomocí algoritmu rankování kandidátů na rozšiřující slova.

Ch. Jin, Ch. Yang: Integrating Hierarchical Feature Selection and Classifier Training for Multi-Label Image Annotation.
V přednášce byl představen dvojúrovňový algoritmus výběru podmnožiny vlastností (feature subset selection) pro snížení dimenzionality prostoru vlastností u vysoce dimenzionálních dat jako např. reprezentace obrázků, na kterých by algoritmus demonstrován.

S. Rendle, Z. Gantner, Ch. Freudenthaler, L. Schmidt-Thieme: Fast Context-aware Recommendations with Factorization Machines.
V přednášce bylo představeno využití tzv. factorization machines (pokrývajících metody faktorizaci matic nebo SVD) pro predikování hodnocení v doporučovacích systémech zahrnujícím kontextové informace (tzv. context-aware rating) s lineární složitostí.

X. Han, L. Sun, J. Zhao: Collective Entity Linking in Web Text: A Graph-Based Method.
V přednášce byla představena metoda propojení názvů vyskytujících se v dokumentu s referenčními entitami ve znalostní bázi se zachycením vzájemných vztahů mezi názvy v dokumentu (tzv. collective entity linking) spočívající ve využití grafové reprezentace vztahů.

Q. Hao, R. Cai, Y. Pang, L. Zhang: From One Tree to a Forest: a Unified Solution for Structured Web Data Extraction.
Přednáška pojednávala o extrakci strukturovaných dat z rozsáhlejších webových stránek, při které jsou naučeny obecné vlastnosti z jedné stránky a aplikovány při extrakci informací ze stránek stejného druhu.

Y. Kim, J. Seo, B. W. Croft: Automatic Boolean Query Suggestion for Professional Search.
Přednáška pojednávala o generování Boolovských dotazů z rozhodovacích stromů, vytvořených z top-k dokumentů základního vyhledání, v prostředí profesionálních vyhledávačů (patentové, medicínské).

H. Wang, H. Huang, F. Nie, Ch. Ding: Cross-Language Web Page Classification via Dual Knowledge Transfer Using Nonnegative Matrix Tri-Factorization.
V přednášce byla představena metoda klasifikace vícejazyčných webových stránek ke stejnému tématu využívající jejich podobnosti založená na učení z různých zdrojů (tzv. transfer learning) a faktorizaci matic.

Y. Yang, Z. Gong, L. Hou U: Identifying Points of Interest by Self-Tuning Clustering.
Přednáška pojednávala o identifikaci bodů zájmu (points of interest, POI) z fotografií geografických míst pomocí self-tuning a reinforcement shlukování s využitím pro vytipování cest.

S. Tatikonda, B. Barla Cambazoglu, F. P. Junqueira: Posting List Intersection on Multicore Architectures.
Přednáška pojednávala o paralelizaci průniku seznamů identifikátorů dokumentů obsahujících daný výraz (tzv. posting list), základní operaci vyhledávačů, rozložením výpočtu na disjunktní nezávislé části zpracovávané paralelně za použití nejjemnější úrovně paralelizace (finest-grained parallelization) na vícejádrových procesorech.

S. Ding, T. Suel: Faster Top-k Document Retrieval Using Block-Max Indexes.
V přednášce bylo představeno vylepšení algoritmů pro operaci top-k zavedením nové indexové blokové struktury (block-max index) umožňující rychlejší operace s posting listy.

Další prezentace (panely, demonstrace, postery)
Workshopů a tutoriálů jsem se neúčastnil, protože moje registrace je nezahrnovala (měly další vlastní poplatky). Z posterů a demonstrací, které se týkaly snad napříč celým spektrem témat konference, mě více zaujaly následující:

Ch. Hoobin, S. Puglisi, J. Zobel: Sample Selection for Dictionary-Based Corpus Compression.
- snížení redundance slovníku u výkonné LZ-like komprimační metody

A. Broschart, R. Schenkel: A Novel Hybrid Index Structure for Efficient Text Retrieval.
- struktura snižující počet předpočítaných indexových seznamů

S. Wang, B.J. Gao, K. Wang, H.W. Lauw: Parallel Learning to Rank for Information Retrieval.
- učení rankovací funkce použitelné pro uspořádání výsledků budoucích dotazů

W.-Y. Lee, L.-Ch. Hsieh, G.-L. Wu, W. Hsu, Y.-F. Su: Multi-Layer Graph-Based Semi-Supervised Learning for Large-Scale Image Datasets Using MapReduce.
- učení na základě grafů vizuální a textové podobnosti obrázků

M. Huggett, E. Rasmussen: The Meta-Dex Suite: Generating and Analyzing Indexes and Meta-Indexes.
- software pro extrakci a indexaci textu z kolekce PDF souborů a vytvoření tzv. meta-indexu

Vlastní prezentace
Neměl jsem vlastní prezentaci.

Shrnutí konference (perspektivní témata apod.)
Na konferenci byly prezentovány průlomové příspěvky z těchto oblastí: reprezentace, analýza a indexování dat (dokumentů) a jejich modely, klasifikace a shlukování, rankování, interaktivní IR (WWW, komunity), analýza dotazů a vyhledávání obsahu, filtrace a ve velké míře doporučování obsahu (perzonalizace). Z příspěvků lze usuzovat, že perspektivními směry dalšího výzkumu jsou: reprezentace, analýza a indexování dokumentů, rankování, analýza dotazů, a zejména vyhledávání a doporučování (převážně webového) obsahu. Na konferenci proběhly demonstrace softwarových nástrojů pro vyhledávání a indexování obsahu a zpracování výsledků vyhledávání a reálných aplikací pro podporu vyhledávání a správu dokumentů celkem 5 firem (Google, eBay, Facebook, Microsoft, Elsevier).

Navázání kontaktů
Shi Yuan (*Baidu*), Max Berenstein (*Elsevier*) - diskuze o možnosti využití data miningu, speciálně formální konceptuální analýzy, v jejich výzkumu

Fotografická dokumentace



Přílohy
Program konference