信息检索与数据挖掘

课程要求（2019）：论文阅读&研讨

文献阅读建议

- **每人阅读一篇文献并做PPT，安排1或2次课讲解（演讲人随机抽取）。**
- **建议从课程内容相关会议的近10年的Best Paper或Honourable Mentions中选取，如**
  - SIGIR (Information Retrieval)
  - WWW (World Wide Web)
  - KDD (Knowledge Discovery and Data Mining)
  - CIKM (Knowledge Management)
    - http://jeffhuang.com/best_paper_awards.html
  - NIPS (Neural Information Processing Systems)
    - https://nips.cc/

Best Paper Awards in Computer Science (since 1996)

- **https://jeffhuang.com/best_paper_awards.html**

- **By**
  Conference: AAAI ACL CHI CIKM CVPR FOCS FSE ICCV ICML ICSE IJCAI INFOCOM KDD MOBICOM NSDI OSDI PLDI PODS S&P SIGCOMM SIGIR SIGMETRICS SIGMOD SODA SOSP STOC UIST VLDB WWW

http://sigir.org/sigir2019/

- The 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval will take place on July 21-25, 2019 in Paris.
  - ACM SIGIR 是国际计算机协会信息检索大会的缩写。SIGIR 专注于信息存储、检索和传播的各个方面，包括研究战略、输出方案和系统评估。
  - 国际信息检索大会的历史可以追溯到1971年。当年，Jack Minker 和Sam Rosenfeld组织召开了ACM SIGIR 的信息存储和检索研讨会。

SIGIR 2018 Best Paper

Should I Follow the Crowd? A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems

The use of IR methodology in the evaluation of recommender systems has become common practice in recent years. **IR metrics** have been found however to be strongly **biased** towards rewarding algorithms that recommend popular …... The fundamental question remains open though **whether popularity is really a bias** we should avoid or not; whether it could be a useful and reliable signal in recommendation, or it may be unfairly rewarded by the experimental biases. …... We build a crowdsourced dataset devoid of the usual biases ……

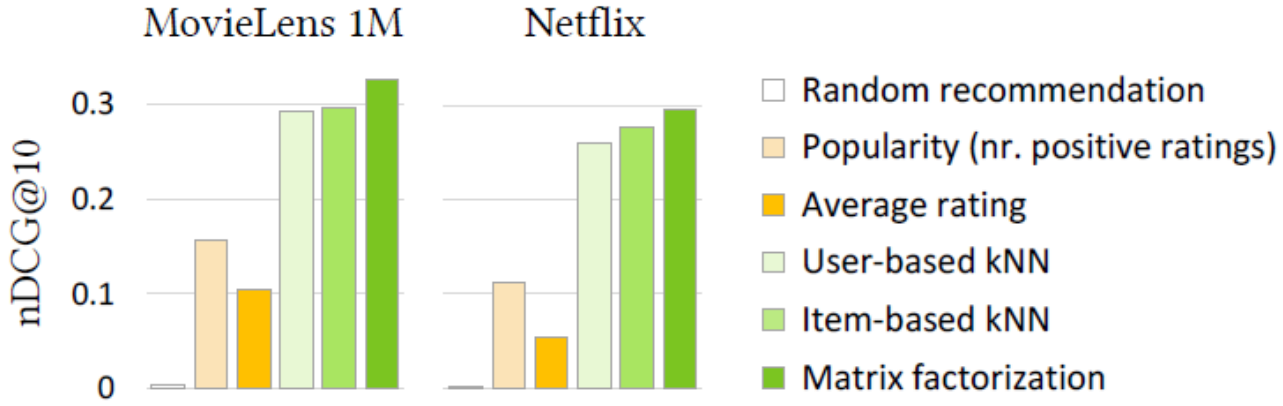

Figure 1: Typical offline experimental results for non-personalized popularity-based recommendation compared to personalized algorithms on two public datasets.

SIGIR 2017 Best Paper

BitFunnel: Revisiting Signatures for Search

- **Since the mid-90s there has been a widely-held belief that signature files are inferior to inverted files for text indexing. In recent years the Bing search engine has developed and deployed an index based on bit-sliced signatures. This index, known as BitFunnel, replaced an existing production system based on an inverted index.……**
- **The BitFunnel algorithm directly addresses four fundamental limitations in bit-sliced block signatures. At the same time, our mapping of the algorithm onto a cluster offers opportunities to avoid other costs associated with signatures. We show these innovations yield a significant efficiency gain versus classic bit-sliced signatures and then compare BitFunnel with Partitioned Elias-Fano Indexes, MG4J, and Lucene.**

https://dl.acm.org/citation.cfm?doid=3077136.3080789

SIGIR 2017 Honourable Mentions(最佳提名)

- **IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models**
  - Jun Wang (University College London), Lantao Yu (Shanghai Jiao Tong University), Weinan Zhang (Shanghai Jiao Tong University), Yu Gong (Alibaba Inc.), Yinghui Xu (Alibaba Inc.), Benyou Wang (Tianjin University), Peng Zhang (Tianjin University), Dell Zhang (Birkbeck, University of London)
- 评价指标设计一直是信息检索技术研究中的核心问题之一，而估计用户的期望收益与期望付出则是搜索用户行为模型的关键组成部分。受模型框架限制，当前几乎所有信息检索评价指标均无法做到同时将用户的期望收益和付出纳入会话终止条件的估计。针对这一问题，计算机系师生受流行电子游戏“**Bejewed**（中文名：宝石迷阵）”机制启发，设计了一个创新性的用户交互模型框架，将期望收益与付出因素重新建模，并把现有的绝大多数评价指标纳入这一框架的范畴。在真实用户行为数据上的实验表明，该框架比现有指标能够更好的预测用户满意程度。

SIGIR 2016 Best Paper
Understanding Information Need: an fMRI Study

- **In this paper, we investigate the connection between an information need and brain activity. Using functional Magnetic Resonance Imaging (fMRI), we measured the brain activity of twenty four participants while they performed a Question Answering (Q/A) Task, where the questions were carefully selected and developed from TREC-8 and TREC 2001 Q/A Track. The results of this experiment revealed a distributed network of brain regions commonly associated with activities related to in-formation need and retrieval and differing brain activity in processing scenarios when participants knew the answer to a given question and when they did not and needed to search.**

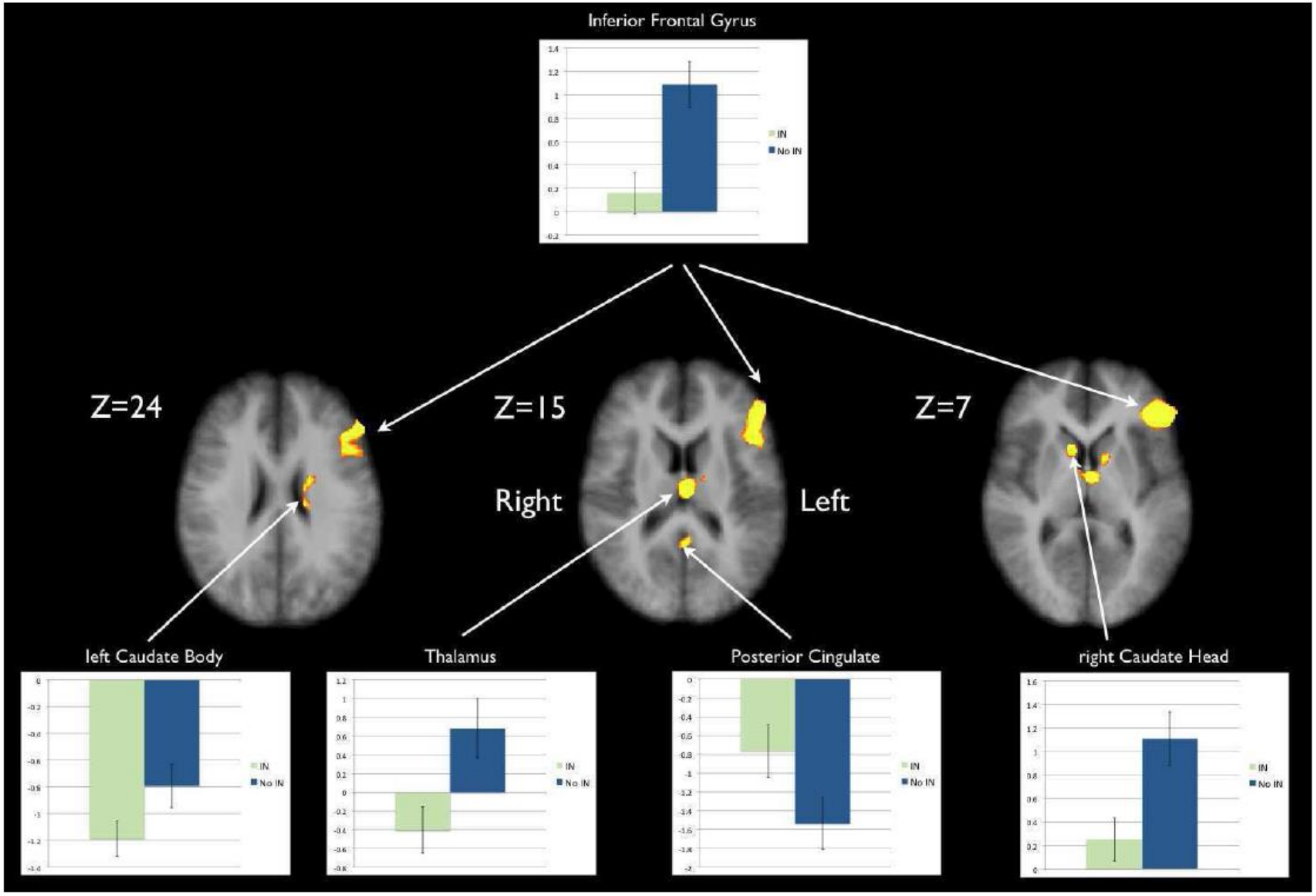

Figure 3: The five activation clusters from Scenario 1 are projected onto the average anatomical structure for three transverse sections. Note that the brains are in radiological format where the left side of the brain is on the right side of the image.

SIGIR 2015 Best Paper

QuickScorer: a Fast Algorithm to Rank Documents with Additive Ensembles of Regression Trees

- **Learning-to-Rank models based on additive ensembles of regression trees have proven to be very effective for ranking query results returned by Web search engines……. Unfortunately, the computational cost of these ranking models is high. ……we present QuickScorer, a new algorithm that adopts a novel bitvector representation of the tree-based ranking model, and performs an interleaved traversal of the ensemble by means of simple logical bitwise operations. ……QuickScorer is able to achieve speedups over the best state-of-the-art baseline ranging from 2x to 6.5x.**

注：线性回归方法可以有效的拟合所有样本点。当数据拥有众多特征并且特征之间关系十分复杂时，构建全局模型的想法一个是困难一个是笨拙。此外，实际中很多问题为非线性的，例如常见到的分段函数，不可能用全局线性模型来进行拟合。 树回归将数据集切分成多份易建模的数据，然后利用线性回归进行建模和拟合。

SIGIR 2014 Best Paper
Partitioned Elias-Fano Indexes

- **The Elias-Fano representation of monotone sequences has been recently applied to the compression of inverted indexes, showing excellent query performance thanks to its efficient random access and search operations. While its space occupancy is competitive with some state-of-the-art methods such as gamma-delta-Golomb codes and PForDelta, it fails to exploit the local clustering that inverted lists usually exhibit, namely the presence of long subsequences of close identifiers. .....We show that our partitioned Elias-Fano indexes offer significantly better compression than plain Elias-Fano, while preserving their query time efficiency. Furthermore, compared with other state-of-the-art compressed encodings, our indexes exhibit the best compression ratio/query time trade-off.**

SIGIR 2013 Best Paper
Beliefs and Biases in Web Search

《互联网搜索中的信仰与偏见(Beliefs and Biases in Web Search)》，作者是来自微软雷蒙德研究院的Ryen White，这也是他第三次获得SIGIR的最佳论文奖(注：前两次分别是2007年和2010年)。这篇文章通过对一系列问卷调查、搜索结果的人工标注以及大规模搜索日志信息的综合分析，探索了预想偏向性(pre-conceived biases)对健康领域搜索的影响。

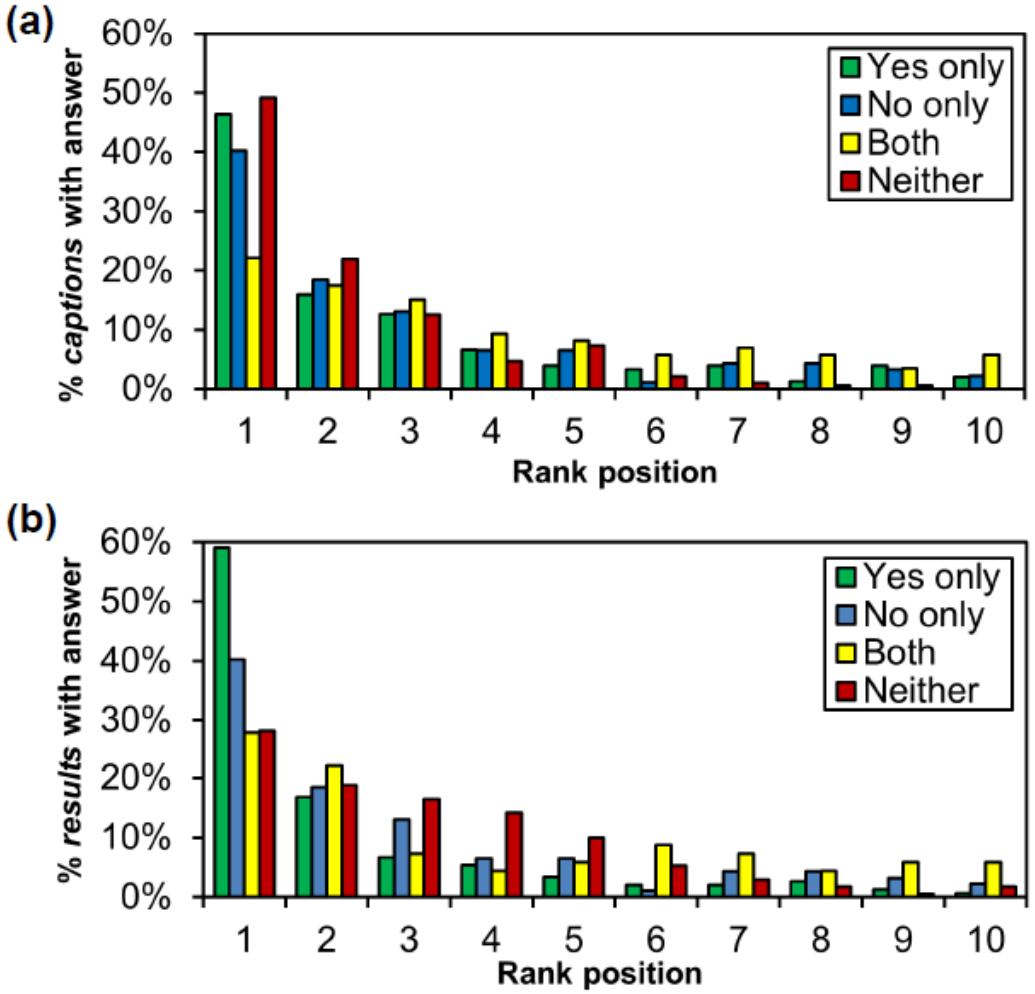

Figure 3. Distribution of *highest-ranked* answers of each type across (a) top 10 captions and (b) top 10 results.

SIGIR 2012 Best Paper
Time-Based Calibration of Effectiveness Measures

- Many current effectiveness measures incorporate simplifying assumptions about user behavior…… In particular, these measures implicitly model users as working down a list of retrieval results, spending equal time assessing each document. In reality, even a careful user, intending to identify as much relevant material as possible, must **spend longer on some documents** than on others. Aspects such as document length, duplicates and summaries all influence the time required. In this paper, we introduce a *time-biased gain* measure, which explicitly accommodates such aspects of the search process.

Best Paper Awards

http://sigir.org/awards/best-paper-awards/

- **2018 Should I Follow the Crowd? A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems**
- **2017 BitFunnel: Revisiting Signatures for Search**
- **2016 Understanding Information Need: an fMRI Study**
- **2015 QuickScorer: A Fast Algorithm to Rank Documents with Additive Ensembles of Regression Trees**
- **2014 Partitioned Elias-Fano indexes**
- **2013 Beliefs and Biases in Web Search**
- **2012 Time-based calibration of effectiveness measures**
- **2011 Find It If You Can: A Game for Modeling Different Types of Web Search Success Using Interaction Data**
- **2010 Assessing the Scenic Route: Measuring the Value of Search Trails in Web Logs**
- **2009 Sources of evidence for vertical selection**
- **2008 Algorithmic Mediation for Collaborative Exploratory Search**

NIPS
https://nips.cc/Conferences/2019/

- The Thirty-third Annual Conference on Neural Information Processing Systems (NIPS) is a multi-track machine learning and computational neuroscience conference that includes invited talks, demonstrations, symposia and oral and poster presentations of refereed papers. Following the conference, there are workshops which provide a less formal setting.
- **Vancouver Convention Center, Vancouver CANADA**
- Sun Dec 8th through Sat the 14th, 2019

NIPS 2018 Best Paper Award

- **Best paper awards:**
  - Non-delusional Q-learning and Value-iteration
  - Optimal Algorithms for Non-Smooth Distributed Optimization in Networks
  - Nearly Tight Sample Complexity Bounds for Learning Mixtures of Gaussians via Sample Compression Schemes
  - Neural Ordinary Differential Equations

https://nips.cc/Conferences/2018/Awards

NIPS 2017 Best Paper Award

- **Best paper awards:**
  - Noam Brown, Tuomas Sandholm. *Safe and Nested Subgame Solving for Imperfect-Information Games*.
  - Hongseok Namkoong, John Duchi. *Variance-based Regularization with Convex Objectives*.
  - Wittawat Jitkrittum, Wenkai Xu, Zoltan Szabo, Kenji Fukumizu, Arthur Gretton. *A Linear-Time Kernel Goodness-of-Fit Test*.
- **Test of time award:**
  - Ali Rahimi, Benjamen Recht. Random Features for Large-Scale Kernel Machines. NIPS 2007.
- **Best Demonstration**
  - Curtis Hawthorne ·Ian Simon ·Adam Roberts ·Jesse Engel · Daniel Smilkov ·Nikhil Thorat ·Douglas Eck. Magenta and deeplearn.js: Real-time Control of DeepGenerative Music Models in the Browser

https://nips.cc/Conferences/2017/Awards

NIPS 2016 Best Paper Award

- **Best paper award:**
  - Aviv Tamar, Yi Wu, Garrett Thomas, Sergey Levine, Pieter Abbeel: *Value Iteration Networks*
- **Best student paper award:**
  - Rong Ge, Jason Lee, Tengyu Ma: *Matrix Completion has No Spurious Local Minimum*
- **Best demonstration: Adam Roberts ·Jesse Engel · Curtis Hawthorne ·Ian Simon ·Elliot Waite · Sageev Oore ·Natasha Jaques ·Cinjon Resnick · Douglas Eck**
  - *Interactive musical improvisation with Magenta*

https://nips.cc/Conferences/2016/Awards

NIPS 2015 Awards

- **Outstanding Demonstration Award**
  - Interactive incremental Question Answering
    - *Jordan Boyd-Graber ·Mohit Iyyer ·He He ·Hal Daumé III*
- **Best papers**
  - Competitive Distribution Estimation: Why is Good-Turing Good
    - *Alon Orlitsky ·Ananda Suresh*
  - Fast Convergence of Regularized Learning in Games
    - *Vasilis Syrgkanis ·Alekh Agarwal ·Haipeng Luo ·Robert Schapire*

https://nips.cc/Conferences/2015/Awards

CIKM Best Paper

ACM International Conference on Information and Knowledge Management (CIKM)

- **2018 Relevance Estimation with Multiple Information Sources on Search Engine Result Pages**
- **2017 Hike: A Hybrid Human-Machine Method for Entity Alignment in Large-Scale Knowledge Bases**
- **2016 Vandalism Detection in Wikidata**
- **2015 Assessing the Impact of Syntactic and Semantic Structures for Answer Passages Reranking**
- **2014 Cross-Device Search**
- **2013 Penguins in Sweaters, or Serendipitous Entity Search on User-generated Content**
- **2012 Gelling, and Melting, Large Graphs by Edge Manipulation**
- **2011 Intent-aware query similarity**
- **2010 MENTA: Inducing Multilingual Taxonomies from Wikipedia**
- **FACeTOR: cost-driven exploration of faceted query results**
- **2009 On the Feasibility of Multi-Site Web Search Engines**
- **2008 Learning to Link with Wikipedia**

SIGKDD Best Papers

- **2018 Adversarial Attacks on Neural Networks for Graph Data**
- **2017 Accelerating Innovation Through Analogy Mining**
- **2016 FRAUDAR: Bounding Graph Fraud in the Face of Camouflage**
- **2015 Efficient Algorithms for Public-Private Social Networks**
- **2014 Reducing the Sampling Complexity of Topic Models**
- **2013 Simple and Deterministic Matrix Sketching**
- **2012 Searching and Mining Trillions of Time Series Subsequences under Dynamic Time Warping**
- **2011 Leakage in Data Mining: Formulation, Detection, and Avoidance**
- **2010 Large linear classification when data cannot fit in memory**
- **Connecting the dots between news articles**
- **2009 Collaborative Filtering with Temporal Dynamics**

SIGKDD is the ACM Special Interest Group (SIG) on Knowledge Discovery and Data Mining

IEEE ICDM Best Paper Awards

IEEE International Conference on Data Mining

- **2018 Discovering Reliable Dependencies from Data: Hardness and Improved Algorithms**
- **2017 TensorCast: Forecasting with Context using Coupled Tensors**
- **2016 KNN Classifier with Self Adjusting Memory for Heterogeneous Concept Drift**
- **2015 Diamond Sampling for Approximate Maximum All-pairs Dot-product (MAD) Search**
- **2014 Ternary Matrix Factorization**
- **2013 Reconstructing Individual Mobility from Smart Card Transactions: A Space Alignment Approach**
- **2012 Scalable Coordinate Descent Approaches to Parallel Matrix Factorization for Recommender Systems**
- **2011 Personalized Travel Package Recommendation**
- **2010 Finding Local Anomalies in Very High Dimensional Space**
- **2009 Explore/Exploit Schemes for Web Content Optimization**
- **2008 Scalable Tensor Decomposition for Multi-Aspect Data Mining**

http://www.cs.uvm.edu/~icdm/Awards/BestPapers.shtml