我的KDD之行

1. **了解KDD**
   ACM SIGKDD 国际会议（简称 KDD）是由 ACM 的数据挖掘及知识发现专委会(SIGKDD)主办的数据挖掘研究领域的顶级年会。它为来自学术界、企业界和政府部门的研究人员和数据挖掘从业者进行学术交流和展示研究成果提供了一个理想场所。本届KDD涵盖了特邀主题演讲（keynote presentations）、论文展板展示(poster sessions）、研讨会（workshops）、短期课程（tutorials）、专题讨论会（panels）、展览（exhibits）、系统演示（demonstrations）、KDD CUP 赛事以及开闭幕式等多项内容。

2. **了解加拿大**
   2017年正直加拿大150周年生日，加拿大的各大景点都开展150周年主题的旅游活动，甚至所有国家公园对所有游人免票，所以今年出国加拿大去班芙国家公园是一个很好的选择。加拿大的官方语言是英语、法语，主要的旅游城市有多伦多、蒙特利尔、卡尔加里等。在多伦多，有个尼日加拉瀑布城是以尼日加拉瀑布为核心的旅游城市，尼亚加拉瀑布(Niagara Falls)位于加拿大安大略省和美国纽约州的交界处，瀑布源头为尼亚加拉河，主瀑布位于加拿大境内，是瀑布的最佳观赏地。

3. **了解哈里法克斯**
   哈利法克斯（Halifax），加拿大新斯科舍省的首府，是加拿大大西洋地区的主要经济中心。本次KDD选在哈利法克斯的世界贸易和会议中心举行。

4. **我的KDD行程**
   8.12 Registration
   8.13 AM 8:00-12:00
   Tutorials：T1 Mining Entity-Relation-Attribute Structures from Massive Text Data
   8.13 PM 1:00-5:00
   Tutorials：T7 Recent Advances in Feature Selection: A Data Perspective
   8.14 AM 8:00-12:00
   Workshop：W1 Mining and Learning from Time Series
   8.14 PM 1:00-5:00
   Workshop：W2: Big Data, IoT Streams and Heterogeneous Source Mining
   8.15 AM 8:00-12:00
   Keynote：Bin Yu ThreePrinciples of Data Science: Predictability,Stability, and Computability
   KDD Exhibit Hall
   8.15 PM 1:00-5:00
   China Chapter Meeting
   8.16 AM 8:00-12:00
   KDD Business Lunch
   8.16 PM 1:00-5:00
   KDD Cup Workshop
   RT8: Representations
   KDD Panel:The Future of Artificially Intelligent Assistants
   8.17 AM 8:30-12:00
   Hands On Tutorial: TensorFlow

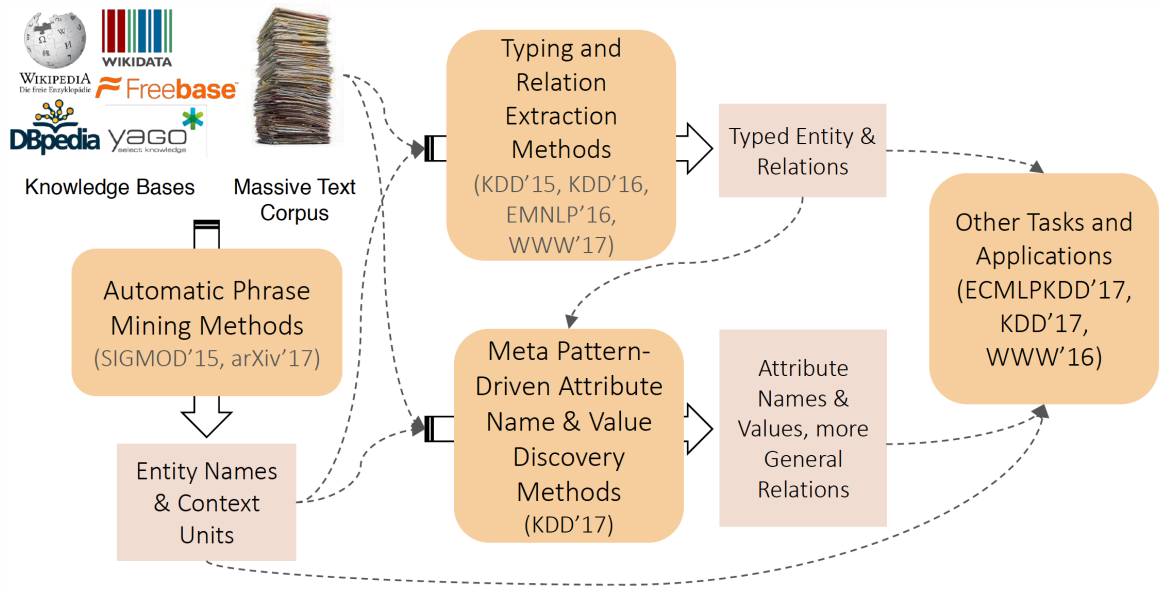
   下面我将选讲Tutorials T1 Mining Entity-Relation-Attribute Structures from Massive Text Data、Hands On Tutorial: TensorFlow以及开幕式上裴健博士的演讲《Pattern Mining Introspection and Prospective》（模式挖掘的回顾与展望）。

5. **Tutorials：T1 Mining Entity-Relation-Attribute Structures from Massive Text Data**

本次教程是从混合文本数据中，挖掘提取实体、关系和属性（ERA）结构。该讲座主要有四个部分，包括：简介、通过短语挖掘进行实体提取、类型化实体和关系、元模式驱动的属性挖掘、从文本中发现结构的应用探索、总结和未来方向。
相关论文：Automatic Entity-Relation-Attribute Structure Mining from Massive Text Data. Jingbo Shang, Xiang Ren, Meng Jiang, Jiawei Han. Computer Science, University of Illinois at Urbana-Champaign. August 11, 2017

5.1 简介

- 从大量文本数据中挖掘结构 由于大约80%的数据是非结构化的文本数据，因此需要建立从大量文本数据中挖掘结构的任务。
- 知识图谱 在知识图谱中，有三个比较重要的概念，分别是实体、关系和属性。
- 结构挖掘
- 一个产品案例：TripAdvisor利用NLP从评论文本中挖掘结构化的因子
- 一个搜索案例：面向集合和实体感知的生物医学文献搜索系统
- 为什么要把文本结构化 结构化的搜索和探索、图挖掘和网络分析、因子分类构造、结构化的特征生成
- 现有技术：利用领域专家知识提取结构
- 本次课程：从混合文本语料中自动挖掘结构 实现各领域应用的快速开发。 提取复杂的结构，而不引入额外的人力。
- 自动的定义： 自动的就是最小化人工参与 仅仅使用存在的通用知识库，而没有任何其他人工参与
- 自动化结构挖掘的方法论



接下来的四个部分将会介绍着四个主要的关键研究成果。

5.2 通过短语挖掘进行实体提取

> - 相关论文
> - Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, Jiawei Han, "Automated Phrase Mining from Massive Text Corpora", submitted to TKDE, under review.
> - Jialu Liu, Jingbo Shang, and Jiawei Han, "Phrase Mining from Massive Text and Its Applications", Synthesis Lectures on Data Mining and Knowledge Discovery, Morgan & Claypool Publishers, 2017.

- Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, Jiawei Han, "Mining Quality Phrases from Massive Text Corpora", in Proc. of 2015 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'15), Melbourne, Australia, May 2015 (**won Grand Prize in Yelp Dataset Challenge, 2015**)
- Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, and Jiawei Han, "Scalable Topical Phrase Mining from Text Corpora", PVLDB 8(3): 305 - 316, 2015. Also, in Proc. 2015 Int. Conf. on Very Large Data Bases (VLDB'15), Kohala Coast, Hawaii, Sept. 2015.

- 定义：优质的短语挖掘 从大规模文档集合中按照质量递减的顺序挖掘短语
- 那么什么样的短语是"优质的"呢

  1. 频繁（Popularity）
  2. 一致（Concordance）
  3. 有信息的（Informativeness）
  4. 完整（Completeness）

- 监督学习方法（语言学分析） 语法树、分块
- 无监督学习方法（统计信号） 限制1：应慎重选择阈值
  限制2：只考虑了满足要求的优质短语的子集
  限制3：以无监督的方式组合不同的信号是困难的
- 半/弱监督学习方法 SegPhrase、AutoPhrase

  5.3 类型化实体和关系

- 相关论文
- Liyuan Liu, Xiang Ren, Qi Zhu, Shi Zhi, Huan Gui, Heng Ji and Jiawei Han, "Heterogeneous Supervision for Relation Extraction: A Representation Learning Approach", in Proc. of 2017 Conf. on Empirical Methods in Natural Language Processing (EMNLP'17), Copenhagen, Denmark, Sept. 2017
- Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare Voss, Heng Ji, Tarek Abdelzaher and Jiawei Han, "CoType: Joint Extraction of Typed Entities and Relations with Knowledge Bases", in Proc. of 2017 World-Wide Web Conf. (WWW'17), Perth, Australia, Apr. 2017.
- Xiang Ren, Wenqi He, Meng Qu, Lifu Huang, Heng Ji, and Jiawei Han, "AFET: Automatic Fine-Grained Entity Typing by Hierarchical Partial-Label Embedding", in Proc. of 2016 Conf. on Empirical Methods in Natural Language Processing (EMNLP'16), Austin, TX, Nov. 2016
- Xiang Ren, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, Jiawei Han, "Label Noise Reduction in Entity Typing by Heterogeneous Partial-Label Embedding", in Proc. of 2016 ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD'16), San Francisco, CA, Aug. 2016
- Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R. Voss, Heng Ji, Jiawei Han, "ClusType: Effective Entity Recognition and Typing by Relation Phrase-Based Clustering", in Proc. of 2015 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'15), Sydney, Australia, Aug. 2015

- 限制：1.一个实体对应多个含义；2.不同实体可能对应同一个含义
- CoType

  5.4 元模式驱动的属性挖掘

- 相关论文
- Meng Jiang, Jingbo Shang, Taylor Cassidy, Xiang Ren, Lance Kaplan, Timothy Hanratty

> and Jiawei Han, “MetaPAD: Meta Patten Discovery from Massive Text Corpora”, in Proc. of 2017 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD’17), Halifax, Nova Scotia, Canada, Aug. 2017

- 定义：属性挖掘

5.5 从文本中发现结构的应用探索

> - 相关论文
> - Huan Gui, Qi Zhu, Liyuan Liu, Aston Zhang, and Jiawei Han, “Expert Finding in Heterogeneous BibliographicNetworks with Locally-trained Embeddings”, submitted for publication
> - Jiaming Shen, Zeqiu Wu, Dongming Lei, Jingbo Shang, Xiang Ren, Jiawei Han, “SetExpan: Corpus-based Set Expansion via Context Feature Selection and Rank Ensemble”, in Proc. of 2017 European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD’17), Skopje, Macedonia, Sept. 2017
> - Meng Qu, Xiang Ren and Jiawei Han, “Automatic Synonym Discovery with Knowledge Bases”, in Proc. of 2017 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD’17), Halifax, Nova Scotia, Canada, Aug. 2017
> - Fangbo Tao, Honglei Zhuang, Chi Wang Yu, Qi Wang, Taylor Cassidy, Lance Kaplan, Clare Voss, Jiawei Han, “Multi-Dimensional, Phrase-Based Summarization in Text Cubes”, Data Eng. Bulletin 39(3), Sept. 2016, pp. 74-84.
> - Jialu Liu, Xiang Ren, Jingbo Shang, Taylor Cassidy, Clare Voss and Jiawei Han, “Representing Documents via Latent Keyphrase Inference”, in Proc. of 2016 Int. World-Wide Web Conf. (WWW’16), Montreal, Canada, April 2016

- 应用1：语义搜索
- 应用2：反面新闻
- 应用3：关键短语筛选
- 应用4：实体扩展

6. **Hands On Tutorial: TensorFlow**

> Github Repository: tensorflow-workshop

在example目录下，包含本次教程的8个文件。

- 00_test_install.ipynb
  测试安装环境，numpy>=1.12.1，jupyter，matplotlib，pandas，Pillow，tensorflow>=1.3.0
- 01_linear_regression_low_level.ipynb
  对模拟数据进行线性回归，并利用TensorBoard进行可视化计算图和相关变量。
- 02_logistic_regression_low_level.ipynb
  对MNIST数据进行逻辑回归，准确率大于90%
- 03_deep_neural_network_low_level.ipynb
  对MNIST数据进行深度全连接网络训练，准确率大于97%
- 04_canned_estimators.ipynb
  对MNIST数据使用Estimators来简化数据、会话、可视化操作，准确率大于97%
- 05_custom_estimators.ipynb
  对MNIST数据使用Estimators的tf.layers来自定义网络结构，准确率大于97.5%
- 06_convolutional_neural_network.ipynb
  对MNIST数据使用Estimators的tf.layers定义CNN网络，准确率大于99%
- 07_structured_data.ipynb
  对UCI的Adult数据使用Datasets API and Estimators，进行DataFrame的结构化数据操作

7. **KDD 2017 Opening Session：Pattern Mining Introspection and Prospective**
   - 前序：啤酒和尿布的故事 什么样的产品组合是顾客频繁一起购买的？ 频繁模式：在数据集中频繁出现的组合 是数据集探索的一个自然的任务
   - 频繁模式的应用 在商店中利用频繁模式对商品布局优化 在web搜索中对于频繁的关键词进行推荐 在化工设计中的频繁子图 在社交网络中的频繁结构 在自然语言处理和理解中的频繁路径
   - Aprioir算法：候选集的生成和测试
   - Apriori算法的提升 主要的观点：1.减少扫描的次数；2.加速候选集的匹配和计数；3.减少候选集的数量 典型的方法：基于hash的技术，事务约简，分隔，采样，动态项目集计数
   - 突破：FP-Growth 数据压缩：聚焦搜索；数据投影：无候选集生成；有约束的频繁模式挖掘
   - 其他的深度优先搜索算法 树投影（C. Aggarwal），垂直形式TID集合遍历（M.J. Zaki）
   - PrefixScan：挖掘序列模式
   - 频繁图模式 挑战：如何有效枚举结构模式？如何有效构建映射数据集？ 在图数据库中频繁图挖掘 在大规模图中生成频繁子图实例
   - 关联图挖掘 On mining Cross-Graph Quasi-Cliques, KDD, 2004
   - 为什么频繁模式挖掘是重要的？ 许多应用：客户分析、推荐、软件bug检测、事件检测、图像和多媒体数据挖掘、化学和生物应用 促进了其他主要的数据挖掘任务：分类、聚类、异常值检测、web挖掘 索引和检索：频繁模式作为特征
   - 利用频繁模式进行分类 如果一个频繁模式X与类C有很强的关联性，那么X->C提供了一个很强的分类能力 CBA（基于分类的关联） CMAR（多分类关联规则） DPClass（判别式基于模式的分类）
   - 利用频繁模式进行聚类 MaPle, ICDM 2003
   - 基于三维模式的聚类 Mining Coherent Gene Clusters from Three-Dimensional Microarray Data, KDD 2004, Best Application Paper Award Runner-up
   - 在Web搜索中频繁模式的应用 Context-Aware Query Suggestion by Mining Click-Through and Session Data, KDD 2008, Best Application Paper Award
   - 颠覆：深度学习 高准确率、高鲁棒性、适合大规模高维度数据集
   - 存在的挑战 普遍性：深度学习要求大量的数据来泛化、解释性、 探索：深度模型擅长回答问题不擅长解决开放性问题
   - 基于模式的思考 模式：数据的总结和概况：一个模式是一个局部模型 很好的解释性；对于探索和思考很好的工具 模式的组合可能生成深度模型：强有力的证据和见解, 深度森林 [Zhi-Hua Zhou et al., IJCAI 2017]
   - 一个根本的挑战 想法：深度模型+模式
   - 模式挖掘和深度模型的结合 是否能够同时取得更高的准确率和更好的解释性？ 想法：从深度模型中学习
     - 将特征空间分为多个局部区域；
     - 估计深度模型非线性的决策边界，通过从一些模式中分段学习线性决策边界
   - 结论 模式和模式挖掘已经被证明是有意义并且实用的 模式挖掘适用于数据 模式挖掘对于数据科学和数据探索是一个基本的工具

关于作者：如有问题请联系 xzhren@pku.edu.cn