网络数据挖掘领域论文的必备知识总结

信息检索和网络数据领域（WWW, SIGIR, CIKM, WSDM, ACL, EMNLP 等）的论文中常用的模型和技术总结

引子：对于这个领域的博士生来说，看懂论文是入行了解大家在做什么的研究基础，通常我们会去看一本书。看一本书固然是好，但是有一个很大的缺点：一本书本身自成体系，所以包含太多东西，很多内容看了，但是实际上却用不到。这虽然不能说是一种浪费，但是却没有把有限力气花在刀口上。

我所处的领域是关于网络数据的处理（国际会议 WWW, SIGIR, CIKM, WSDM, ACL, EMNLP,等）

我列了一个我自己认为的在我们这个领域常常遇到的模型或者技术的列表，希望对大家节省时间有所帮助：

1. 概率论初步

主要常用到如下概念：初等概率定义的三个条件，全概率公式，贝叶斯公式，链式法则，常用概率分布（Dirichlet 分布，高斯分布，多项式分布，泊松分布 m）

虽然概率论的内容很多，但是在实际中用到的其实主要就是上述的几个概念。基于测度论的高等概率论，几大会议（www，sigir 等等）中出现的论文中基本都不会出现。

2. 信息论基础

主要常用的概念：熵，条件熵，KL 散度，以及这三者之间的关系，最大熵原理，信息增益(information gain)

3. 分类

朴素贝叶斯，KNN，支持向量机，最大熵模型，决策树的基本原理，以及优缺点，知道常用的软件包

4. 聚类

非层次聚类的 K-means 算法，层次聚类的类型及其区别，以及算距离的方法（如 single，complete 的区别 a），知道常用的软件包

5. EM 算法

理解不完全数据的推断的困难，理解 EM 原理和推理过程

6. 蒙特卡洛算法（特别是 Gibbs 采样算法）

知道蒙特卡洛算法的基本原理，特别了解 Gibbs 算法的采样过程；Markov 随机过程和 Markov chain 等

7. 图模型

图模型最近几年非常的热，也非常重要，因为它能把之前的很多研究都包括在内，同时具有直观之意义。如 CRF, HMM，topic model 都是图模型的应用和特例。

a.了解图模型的一般表示（有向图和无向图模型 x），通用的学习算法（learning）和推断算法（inference），如 Sum-product 算法，传播算法等

b. 熟悉 HMM 模型，包括它的假设条件，以及前向和后向算法；

c. 熟悉 LDA 模型，包括它的图模型表示 i，以及它的 Gibbs 推理算法；变分推断算法不要求掌握。

d.了解 CRF 模型，主要是了解它的图模型表示，如果有时间和兴趣 a，可以了解推理算法；

e. 理解 HMM,LDA, CRF 和图模型的一般表示，通用学习算法和推理算法之间的联系和差别；

f.了解 Markov logic network（MLN），这是建构在图模型和一阶逻辑基础上的一种语言，可以用来描述很多现实问题，初步的了解，可以帮助理解图模型；

8. topic model

这个模型的思想被广泛地应用，全看完没有必有也没有时间，推荐如下：

a.深入理解 pLSA 和 LDA，同时理解 pLSA 和 LDA 之间的联系和区别；这两个模型理解后，大部分的 topic model 的论文都是可以理解的了，特别是应用到 NLP 上的 topic model。同时，也可以自己设计自己需要的非层次 topic model 了。

b.如果想继续深入，继续理解 hLDA 模型，特别是理解背后的数学原理 Dirichlet Process，这样你就可以自己设计层次 topic model 了;

c.对于有监督的 topic model，一定要理解 s-LDA 和 LLDA 两个模型，这两个模型体现了完全不同的设计思想，可以细细体会，然后自己设计自己需要的 topic model；

d. 对于这些模型的理解，Gibbs 采样算法是绕不开的坎；

9. 最优化和随机过程

a.理解约束条件是等号的最优化问题及其 lagrange 乘子法求解；

b.理解约束条件是不等号的凸优化问题，理解单纯形法；

c.理解梯度下降法，模拟退火算法；

d.理解爬山法等最优化求解的思想

e.随机过程需要了解随机游走，排队论等基本随机过程（论文中偶尔会有，但不是太常见 n），理解 Markov 随机过程（非常重要，采样理论中常用 l）；

10. 贝叶斯学习

目前越来越多的方法或模型采用贝叶斯学派的思想来处理数据，因此了解相关的内容非常必要。

a. 理解贝叶斯学派和统计学派的在思想和原理上的差别和联系；

b. 理解损失函数，及其在贝叶斯学习中的作用；记住常用的损失函数；

c. 理解贝叶斯先验的概念和四种常用的选取贝叶斯先验的方法；

d. 理解参数和超参数的概念，以及区别；

e. 通过 LDA 的先验选取（或者其它模型 i）来理解贝叶斯数据处理的思想；

11. 信息检索模型和工具

a.理解常用的检索模型；

b.了解常用的开源工具（lemur，lucene 等 ng）

12. 模型选择和特征选取

a.理解常用的特征选择方法，从而选择有效特征来训练模型;

b.看几个模型选择的例子，理解如何选择一个合适模型；（这玩意只能通过例子来意会了）