

分享我看过的资料给你

专栏主体内容已经结束了，在专栏写作过程中，我阅读了很多业界公开的资料，我觉得有必要整理出来，供想深入阅读的人继续去找虐。

整体来说，在选择参考文献时，我偏爱那些由公司发表的。因为推荐系统本质上还是一种非常依赖实践的算法应用方向，并且，这些商业公司论文中的技术内容也在他们实际的场景中经过了检验。

另外，更多的内容是来自我自己的大脑中，所以我在下面列出来的只是一部分，在经过反复删减之后，保留了这些，有中文有英文，一般来说英文居多。有较理论化的，如优化理论，更多的是较实践派，可以学完即用。这些资料分成这么几个类型。

1. 论文：以论文形式发表的，期刊数据库中可以下载到。
2. 网络文章：就是在网上自由流传的内容或者博客，为了方便阅读，我将它们保存为PDF格式。
3. 演示文稿：就是作者曾公开演讲过的内容，相对来说不是那么严谨，但是更容易理解。
4. 书：推荐系统相关的书较少，我在专栏中参考过的书只有一本（附件中不提供书的电子文档）。

以上的参考文献我按照章节顺序列在了下面，我还在后面附上一个推荐书单。你可以点击查看。

原理篇

内容推荐

- 题目：Bag of Tricks for Efficient Text Classification
- 类型：论文
- 作者：Facebook
- 说明：

Facebook开源的文本处理工具fastText背后原理。可以训练词嵌入向量，文本多分类，效率和线性模型一样，效果和深度学习一样，值得拥有。

- 文献：The Learning Behind Gmail Priority Inbox
- 类型：论文
- 作者：Google
- 说明：

介绍了一种基于文本和行为给用户建模的思路，是信息流推荐的早期探索，Gmail智能邮箱背后的原理。

- 题目：Recommender Systems Handbook(第三章，第九章)
- 类型：书
- 作者：Francesco Ricci等
- 说明：

这本书收录了推荐系统很多经典论文，话题涵盖非常广，第三章专门讲内容推荐的基本原理，第九章是一个具体的基于内容推荐系统的案例。

- 题目：文本上的算法
- 类型：网络文章(网络免费版，已有成书《文本上的算法:深入浅出自然语言处理》，内容更丰富)
- 作者：路彦雄
- 说明：

介绍了文本挖掘中常用的算法，及基础概念。内容涉及概率论，信息论，文本分类，聚类，深度学习，推荐系统等。

- 题目：LDA数学八卦
- 类型：网络文章
- 作者：Rickjin(@靳志辉)
- 说明：

由浅入深地讲解LDA原理，对于实际LDA工具的使用有非常大的帮助。

近邻推荐

- 题目：Amazon.com recommendations: item-to-item collaborative filtering
- 类型：论文
- 作者：Amazon
- 说明：

介绍Amazon的推荐系统原理，主要是介绍Item-Based协同过滤算法。

- 题目：Slope One Predictors for Online Rating-Based Collaborative Filtering
- 类型：论文
- 作者：Daniel Lemire等
- 说明：

Slope One算法。

- 题目：Item-Based Collaborative Filtering Recommendation Algorithms
- 类型：论文
- 作者：Badrul Sarwar等

- 说明：

GroupLens的研究团队对比了不同的Item-to-Item的推荐算法。

- 题目：Collaborative Recommendations Using Item-to-Item Similarity Mappings
- 类型：专利
- 作者：Amazon
- 说明：

是的，Amazon申请了Item-Based算法的专利，所以如果在美上市企业，小心用这个算法。

- 题目：Recommender Systems Handbook（第4章）
- 类型：书
- 作者：Francesco Ricci等
- 说明：

第四章综述性地讲了近邻推荐，也就是基础协同过滤算法。

矩阵分解

- 题目：Matrix Factorization and Collaborative Filtering
- 类型：演示文稿
- 作者：Daryl Lim
- 说明：

从PCA这种传统的数据降维方法讲起，综述了矩阵分解和协同过滤算法。矩阵分解也是一种降维方法。

- 题目：Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model
- 类型：论文
- 作者：Yehuda Koren
- 说明：

把矩阵分解和近邻模型融合在一起。

- 题目：BPR- Bayesian Personalized Ranking from Implicit Feedback
- 类型：论文
- 作者：Steffen Rendle等
- 说明：

更关注推荐结果的排序好坏，而不是评分预测精度，那么BPR模型可能是首选，本篇是出处。

- 题目：Collaborative Filtering for Implicit Feedback Datasets

- 类型：论文
- 作者：Yifan Hu等
- 说明：

不同于通常矩阵分解处理的都是评分数据这样的显式反馈，本文介绍一种处理点击等隐式反馈数据的矩阵分解模型。

- 题目：Matrix Factorization Techniques For Recommender Systems
- 类型：论文
- 作者：Yehuda Koren等
- 说明：

本文是大神Yehuda Koren对矩阵分解在推荐系统中的应用做的一个普及性介绍，值得一读。

- 题目：The BellKor Solution to the Netflix Grand Prize
- 类型：论文
- 作者：Yehuda Koren
- 说明：

也是一篇综述，或者说教程，针对Netflix Prize的。

模型融合

- 题目：Adaptive Bound Optimization for Online Convex Optimization
- 类型：论文
- 作者：Google
- 说明：

FTRL是CTR预估常用的优化算法，本文介绍FTRL算法原理。

- 题目：在线最优化求解
- 类型：网络文章
- 作者：冯扬
- 说明：

是对FTRL的通俗版解说。

- 题目：Ad Click Prediction: a View from the Trenches
- 类型：论文
- 作者：Google
- 说明：

FTRL工程实现解读。

- 题目：Factorization Machines

- 类型：论文
- 作者：Steffen Rendle
- 说明：

提出FM模型的论文，FM用于CTR预估。

- 题目：Field-aware Factorization Machines for CTR Prediction
- 类型：论文
- 作者：Yuchin Juan
- 说明：

FFM模型，用于CTR预估。

- 题目：Practical Lessons from Predicting Clicks on Ads at Facebook
- 类型：论文
- 作者：Facebook
- 说明：

提出了LR + GBDT的CTR预估模型

- 题目：Wide & Deep Learning for Recommender Systems
- 类型：论文
- 作者：Google
- 说明：

提出融合深度和宽度模型的Wide&Deep模型，用于CTR预估。

Bandit算法

- 题目：Introduction to Bandits– Algorithms and Theory Part 1– Bandits with small sets of actions
- 类型：演示文稿
- 作者：Jean-Yves Audibert 等
- 说明：

介绍bandit算法概念，理论和算法，这部分主要针对小的选项候选集。

- 题目：Introduction to Bandits– Algorithms and Theory Part 2– Bandits with large sets of actions
- 类型：演示文稿
- 作者：Jean-Yves Audibert 等
- 说明：

介绍Bandit算法概念，理论和算法，这部分主要针对较大的选项候选集。

- 题目：A Contextual-Bandit Approach to Personalized News Article Recommendation
- 类型：论文
- 作者：Yahoo
- 说明：

Linucb的原始论文，考虑上下文的Bandit算法。

- 题目：Collaborative Filtering Bandits
- 类型：论文
- 作者：Shuai Li等
- 说明：

Bandit 算法与协同过滤结合，提出COFIBA算法。

深度学习

- 题目：Deep Neural Networks for YouTube Recommendations
- 类型：论文
- 作者：Google
- 说明：

介绍YouTube视频推荐系统在深度神经网络上的尝试。能从中看到wide&deep模型的影子。

- 题目：Efficient Estimation of Word Representations in Vector Space
- 类型：论文
- 作者：Google
- 说明：

Word2Vec的作者在这篇文章中提出了一种词嵌入向量学习方法，也就是把开源工具包Word2Vec背后的模型详细介绍了一次。理论上很简单，更多是一些工程技巧的分享。Word2Vec给推荐系统带来了一种新的隐因子向量学习方法，深陷评分预测泥潭的矩阵分解被开拓了思路。

- 题目：Item2Vec: Neural Item Embedding for Collaborative Filtering
- 类型：论文
- 作者：Microsoft
- 说明：

这篇就是借鉴了word2vec在语言建模中的思路，为推荐系统的行为建模，从中为物品学习嵌入向量。

- 题目：Learning Representations of Text using Neural Networks
- 类型：演示文稿
- 作者：Google

- 说明：

理解为word2vec作者写一个教程。

- 题目：Long Short-Term Memory
- 类型：论文
- 作者：Sepp Hochreiter等
- 说明：

可以用来为序列建模的LSTM，实际上在1997年就发表论文了，只是在10几年后才大火。

- 题目：An Empirical Exploration of Recurrent Network Architectures
- 类型：论文
- 作者：Google
- 说明：

Google在RNN模型使用上的经验分享。

- 题目：Recurrent Neural Networks for Collaborative Filtering
- 类型：网络文章
- 作者：Erik Bernhardsson
- 说明：

这是Erik Bernhardsson在Spotify期间所做的尝试，用RNN自动构建音乐播单。Erik Bernhardsson还有一项开源项目Annoy，用于稠密向量的近邻搜索，在推荐系统中也用得较多。

其他实用算法

- 题目：Detecting Near-Duplicates for Web Crawling
- 类型：论文
- 作者：Google
- 说明：

在这篇论文中提出了simhash算法，用于大规模网页去重。

- 题目：Weighted Random Sampling over Data Streams
- 类型：论文
- 作者：Pavlos S. Efraimidis
- 说明：

对流式数据的加权采样。

- 题目：Weighted Sampling Without Replacement from Data Streams
- 类型：论文：

- 作者：Vladimir Braverman等
- 说明：

介绍了两种对流式数据的加权采样

工程篇

常见架构

- 题目：Activity Feeds Architecture
- 类型：演示文稿
- 作者：Etsy
- 说明：

本文非常详细地介绍了社交动态信息流的架构设计细节。

- 题目：Atom Activity Streams 1.0
- 类型：规范文档
- 作者：Activity Streams Working Group
- 说明：

这是一份动态信息流数据模型的协议规范文档，由Activity Streams Working Group共同发出，这个组织包含Google和Microsoft。

- 题目：Beyond the 5 stars (Netflix Recommendations)
- 类型：网络文章
- 作者：Netflix
- 说明：

Netflix详细宏观上介绍了自家推荐系统的产品形态，不只是比赛中的评分预测那么简单的。

- 题目：System Architectures for Personalization and Recommendation
- 类型：网络文章
- 作者：Netflix
- 说明：

Netflix 推荐系统的架构介绍。

- 题目：Information Seeking–Convergence of Search, Recommendations and Advertising
- 类型：论文
- 作者：H Garcia–Molina等
- 说明：

探讨搜索、推荐、广告三者架构统一。

关键模块

- 题目：Overlapping Experiment Infrastructure– More, Better, Faster Experimentation
- 类型：论文
- 作者：Google
- 说明：

ABTest实验平台的扛鼎之作，Google出品，值得拥有。

- 题目：TencentRec: Real-time Stream Recommendation in Practice
- 类型：论文
- 作者：腾讯
- 说明：

介绍了腾讯内部的实时推荐系统架构。

- 题目：Personalization at Spotify using Cassandra
- 类型：网络文章
- 作者：Spotify
- 说明：

介绍了Spotify在推荐系统所用到的数据存储中间件。

效果保证

- 题目：Tutorial on Robustness of Recommender Systems
- 类型：演示文稿
- 作者：Neil Hurley
- 说明：

本文非常详细讨论了对推荐系统的攻击和防护，并有实验模拟。

- 题目：Recommender Systems Handbook(第八章)
- 类型：书
- 作者：Francesco Ricci等
- 说明：

该书第八章介绍了能见到的几乎所有推荐系统评价指标，只是实际上用不到这么多指标。

其他书目

1. Pattern Recognition and Machine Learning（机器学习基础，有此一本足够了）
2. 推荐系统实践（国内唯一一本非翻译的推荐系统书籍，入门必选）
3. 信号与噪声（介绍贝叶斯统计的一本科普书）

4. 复杂（推荐系统面对的是复杂网络，了解复杂系统和复杂网络的特点，有助于开脑洞）
5. 信息简史（既然是信息经济，当然要读一本关于信息的历史）

知道你们不会读的，所以就不推荐太多了。

但愿我这个激将法有助于你学习进步。



推荐系统36式参考资...选集.zip
35.9MB