

哈哈，成功啦~

Proj 2 Proposal

刘斌 熊世裕 李书昂

2014013466 2014013413 2014013431

(1) 文献阅读

《A Contextual-Bandit Approach to Personalized News Article Recommendation》

Main idea:

这篇论文将新闻文章推荐系统建立为一个上下文相关的系统，其中学习算法根据用户和文章的相关信息自动推荐文章给用户。该系统采用了基于用户点击反馈从而最大化总用户点击量的文章选择技巧。

这篇文章的系统已经成功应用于 Yahoo! Front Page Today Module，结果显示，12.5%的点击提升与标准上下文算法相比，而且数据变少时，优势更加明显。

本文实现一个简单而且可靠的方法直接从记录的事件中评估上下文算法，从而可以避免建立同步过程中的问题。而且新的算法 LinUCB 显示当数据集较为稀疏时，即使数据规模很大，正确率也非常高。最后，用户的兴趣会根据时间变化，所以考虑算法的暂时信息非常重要。

《Content-Based Book Recommending Using Learning for Text Categorization》

Main idea:

许多推荐系统都是基于用户的选择，使用协同过滤算法来进行推荐。但是，对于尚未被任何用户评价过的新物品，协同过滤算法不能给出较好的推荐结果。相反，基于内容的推荐算法此时就能够发挥作用。这篇论文提出了一种基于内容的书籍推荐系统，利用信息提取技术和机器学习算法对文本进行分类。下面是这个系统的详细描述：

- 1) 信息提取，创建数据库。
- 2) 学习一个分析器。该过程使用贝叶斯分类和多项式文本模型，依据用户对已阅读的书籍的评分来进行学习一个分类器。
- 3) 推荐并同时根据分析器提供推荐理由。

总结：对于数字图书馆来说，基于用户的兴趣而不是特定的需求推荐书籍和其他信息资源是其提供的服务中一个重要的部分。不像协同过滤算法，基于内容的推荐对于那些未被评分过的物品有着更加杰出的表现。也能够针对用户独特的喜好来进行高质量的、个性化的推荐。本文所构建的系统是一个初级的基于内容的推荐书籍推荐系

统，使用了简单的贝叶斯学习算法和从网络上提取出来的关于书籍的信息。一些简单的实验表明这种方法在缺失用户之间的信息的条件下提供了精确的推荐。

对于那些已经被许多人评价过的物品，协同过滤算法能提供很精确的推荐。但是对于那些尚未被评价的物品，基于内容的推荐系统有着完美的表现。因此，如果将两种方法联合起来，将会行成一个更好更完美的推荐系统。

《Collaborative Filtering: A Machine Learning Perspective》

Main idea:

这是一本对推荐系统介绍较为完整的经典书籍。其中提到、解释、并实验了推荐系统中可能用到的几乎所有算法。其中包括：

- 1) 推荐系统的相关介绍
- 2) 推荐系统的正式定义、公式化和相关工作
- 3) 推荐系统所需要的一些背景知识。
- 4) 推荐系统中需要使用到的一些分类和回归算法的相关讨论和实验以及结果。
- 5) 推荐系统中所需要使用到的聚类算法，包括标准聚类和分层聚类的预测、复杂度和实验结果
- 6) 维度约简，这部分包括两个子内容：奇异值分解和主成分分析
- 7) 基于概率的评分模型
 - a) 多项式模型 (The Multinomial Model)
 - b) 混合多项式模型 (Mixture of Multinomial Model)
 - c) 视面模型 (The Aspect Model)
 - d) 用户评分分析模型 (The User Rating Profile Model)
 - e) 分布模型 (The Attribute Model)
- 8) 上述提到所有模型和方法的比较。

通过以上内容的介绍，作者总结入下：

本书以较易理解的方式讲解了基于评分的，纯净的，时序的协同过滤算法。并对许多方法进行了详细的讨论和推导，展示了他们和标准机器学习算法的关系，分析了他们的时间、空间复杂度。

这本书展示了 original GroupLens 方法是著名的 K-最近邻分类的变种，我们介绍了标准朴素贝叶斯的一个新的应用，我们讨论了一些基于用于聚类的评分预测方法

并介绍了一个基于标准 K-中位数的新的聚类方法。我们复习了维度约简技术，包括奇异值分解，权值奇异值分解，主成分分析，概率主成分分析，和因子分析。对于权值奇异值分解，我们介绍了一个新的评分预测技术。在概率模型中我们复习了多项式模型，混合多项式模型，视面模型，和用户评分模型。我们也介绍了分布模型家族，这对于协同过滤来说是一个全新的产生模型。

这些研究的主要结果是新的和重要的评级预测方法的识别。通过比较发现，K-平均聚类方法被发现具有优秀的评级预测能力。由于其输入相对简单，因此它是众多方法中能较好的运用于交互式推荐的一个推荐算法。分布模型家族在设计时就有很强的直观性、有吸引人的泛化能力。虽然二值分布模型在学习方面和预测方面都显得特别复杂，但是通过我们的研究发现，它达到最佳的整体预测精度。

《ARSYS - Article Recommender System》

main idea:

本篇论文介绍了一个名为 ARSYS 的文章推荐系统，这个系统融合了一般推荐系统常用的两种实现方式：基于内容(content based)以及协同筛选(collaborative filtering)，并且加入了基于元数据的推荐。用户可在这三种方式中选择最合适自己的进行推荐。

基于内容的实现方式通过文章的关键词去找到相似的文章；协同筛选通过 P2P 网络实现；最后一种方式基于存放有文章信息的语义库实现，支持自然语言的查询。

P2P(Peer to Peer)网络为需要进行大数据共享、内容分发、应用水平上的多点传送的APP提供了一个优秀的平台。它的底层是由DHT (Distributed Hash Table) 实现的。

用户在查询时的问题会被写成 ARSY-QL 语言形式，这是一种类似于 SQL 的语言，但更接近于自然语言。此外，在通过语义进行查询时，系统支持用户使用他们的 ARSY-QL 语言对查询进行精确。

结论：论文介绍了一个融合了多种推荐方式，实现了类似自然语言查询，并且利用 P2P 实现了高效率计算的推荐系统: ARSYS。对于进一步的提高，作者考虑对不同用户的进程实现平行操作，在多线程方面做出优化。

《Recommender systems survey》

main idea:

本篇论文对推荐系统进行了较为详细的介绍，并对实现推荐系统的方式之一协同筛选和它的算法进行了介绍。论文介绍了他们的演变过程，并给出了他们的原始分类。

推荐系统和网络是同步发展的，它的变革陪伴着网络的变革。

第一代推荐系统通过传统的网络从这三个来源搜集数据：1、基于内容的数据 2、人口统计的数据 3、基于操作历史的数据；

第二代推荐系统伴随着第二代网络，它可以搜集社交信息（你的朋友、粉丝、关注的人等等）；

第三代推荐系统通过融合在因特网上的设备提供的信息去进行操作，一些外置的信息将被纳入（比如一些从设备上采集到的你的身体信息）。

未来的在推荐系统上的工作将集中在优化现有的方式和算法上，提高推荐结果的质量。同时，新的一些领域有：1、对现有的使用不同类型信息的推荐算法进行结合 2、实现推荐系统进程的安全性和保密性 3、设计灵活的结构去实现不同类型数据的自动分析……

(2) 我们的推荐系统实现思路

我们计划选用基于领域的协同过滤算法来实现推荐系统，详细说明如下：

首先，我们将实现

- 1、 基于用户过去的行为给用户推荐与用户读过的文章相似的文章。
- 2、 基于相似用户的行为对用户进行推荐。即推荐与该用户具有相同兴趣的用户所阅读过的书籍给该用户。
- 3、 按照一定的权重将上述两个推荐列表合并。不断调整上述两个模块返回结果的权重，以获取一个较好的推荐结果。

其次，如果时间和精力足够，我们计划再实现基于物体特征的推荐系统。学习每一个用户和每一本书的特征，然后依据这些特征给用户进行推荐。然后将这个模块加入之前的系统中，并调整合适的权重。

最后，我们将采取多方面的指标来进行离线测试，包括准确度 precision，召回率 recall，覆盖率 cover，F 值，AP 值等。