社会化推荐在人人网的应用

作者：张叶银

阅读数：5897 2012 年 2 月 27 日 话题：《架构师》月刊 架构

1 推荐引擎简介

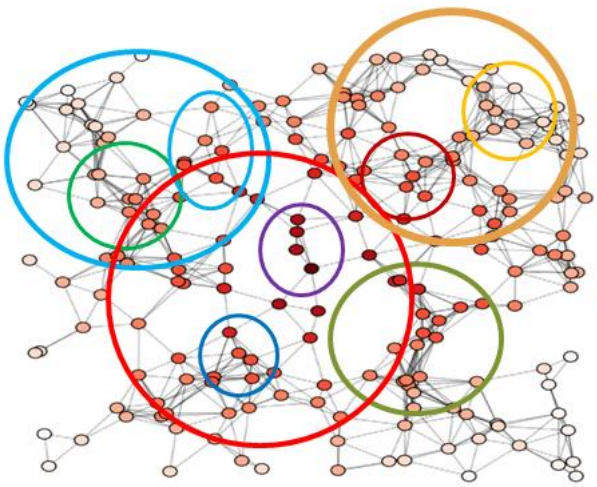
互联网快速发展的今天，数据爆发式的增长，使得用户在浏览网页时不得不花费大量时间用以筛选目标信息。推荐引擎的核心使命就在于，采用数据挖掘和机器学习方法，替用户发现自己感兴趣的事物，迅速定位用户真实所需。推荐引擎随着电子商务的蓬勃发展越来越受到业界的关注，再加上 web2.0 的兴起，推荐引擎在 SNS 领域也越来越发挥出巨大的潜力。从用户的角度来讲，面对各种信息源，以及层出不穷的社交网站使得用户自主获取资讯方式非常低效；从企业角度来讲，互联网高速发展的十年中，已积攒了大量数据，合理挖掘这些大量数据的价值，已成为互联网公司提高产品渗透率或盈利的有效途径。

推荐引擎通过数据挖掘和机器学习的方法，分析用户资料和历史行为，定位用户的兴趣爱好，将用户可能感兴趣的物品或内容推荐给用户。常用的协同过滤推荐算法是通过分析用户的历史记录，度量用户与用户，物品与物品之间的相似性，从而找到用户感兴趣的物品或相似性较高的物品。这项技术已广泛应用于商业推荐系统中，例如电子商务网站亚马逊的物品推荐，视频网站 YouTube 的视频推荐等。与传统的推荐引擎不同，社会化推荐考虑了人与人之间的关系，通过真实的人际关系，提高推荐结果的精准度。例如，好友推荐的目的是帮助用户找到好友，迅速构建用户社交图谱，形成可持续发展的网络生态圈。社会化推荐对传统的推荐引擎提出了新的需求和挑战，具体体现在其个性化和复杂性上。其中个性化是指推荐引擎能够提供用户真实所需以及被用户信赖的推荐结果，复杂性是指结构化大数据的存储和计算对推荐引擎性能提出更高的要求。

2 人人网社会化推荐框架

2.1 社交图谱的构建

人人网是一家发展迅速的实名制社交网站，社交图谱（Social Graph）是人人网的基础，它是真实人际关系的映射。通过社交网络服务，用户可以找到旧时的好友，以及拓展自己新的人际关系。其中，推荐引擎一方面需要帮助用户迅速建立自己的社交圈子，另一方面要给用户提供优质和个性化的资讯和内容。这些应用都离不开对社交图谱的分析和挖掘，就人人网来说，其社交图谱的分析从以下三方面着手：



1. **团体发现**

系属性。例如，用户的好友可以分为同学、亲属、同事等团体。与之相对应，用户簇则是对用户集合进行无监督的分簇，发现用户集合内在的层次关系。例如将同一学校的用户分成以班级为单位的簇。团体发现可以应用于好友推荐、用户隐私控制、新鲜事分组等。

2. **亲密度**

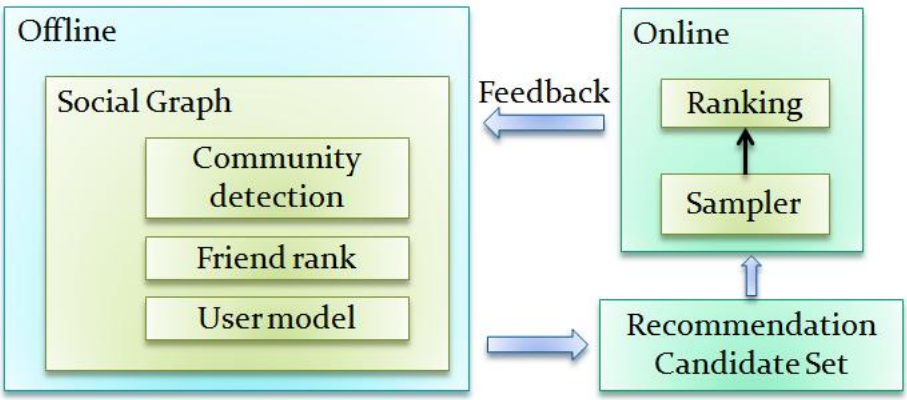
   亲密度用于度量两个用户之间的距离，对应于社交图谱中边的权值或用户之间最短加权路径。其目的是反映现实生活中用户之间的亲密度。亲密度可以通过分析用户的行为，特别是用户之间的交互行为来获取。从社交图谱中直接计算任意两点间的距离非常困难，其时间复杂度随着图中节点个数增加呈指数型增长。所以在实际计算过程中，一般采取定界或嵌入的方法来计算节点间的最短路径。

3. **用户模型**

   用户模型是网站提供个性化服务的基础。用户模型的建立是一个复杂的过程。首先，它需要分析大量的用户数据。这些数据来自不同的域，域与域之间的差异导致很难制定统一的量化方法。例如，不同阶段的用户在网站上的行为表现不同，分析的方法也相应有所区别。其次，需要建立用户特征之间的关系。用户的特征是多维度的，并且现实中一些人口学特征和用户行为具备一定程度上的因果关系。不同年龄用户的兴趣和偏好有所不同，这就需要去发现不同因素间的依赖关系。贝叶斯理论就提供了一个很好的理论框架，经常被用在分类决策中。然而，与传统的模型（例如 Latent DirichletAllocation）相比，用户模型的结构则更为复杂，相应的学习代价更高。

2.2 社会化过滤方法的应用

早期的好友推荐系统只考虑用户与二度好友的共同好友数目。共同好友数目越多，被推荐的可能性就越高。二度好友即好友的好友，对新用户来说，与其二度好友的共同好友数目均较少，故推荐好友的精准性相对较低。此外，一般用户的二度好友数目比较庞大，仅依据共同好友数目很难发现用户的社交关系层次。基于社交图谱，人人网所采用的社会化推荐引擎在 SNS 中的应用效果显著，其总体框架如下图所示。



从图中看到，我们所采用的好友推荐是多种推荐策略融合的结果。具体来讲，主要体现在以下三个方面：

1. **候选集生成**

   为了弥补通过二度好友的共同好友数目推荐好友的不足，我们采用好友聚类的方法来更加准确定位用户的社交圈子，并将圈子内的好友推荐给用户。这一方法的核心思想是通过强关系找到弱关系。强关系是指社交图谱中联系非常紧密的用户集合。我们可以通过图中边的紧致程度寻找强关系，例如最大的完全子图中每两个顶点之间都有边。聚类的目的是以这些强关系为出发点，找出弱关系，将弱关系推荐给用户。聚类需要度量用户与用户之间的亲密度。常用的方法是统计用户浏览行为，认为用户之间的交互越频繁，他们之间的亲密度越高。由于记录的用户浏览行为种类多，每种行为的权重各不相同，容易导致距离的度量失衡，而强关系的利用可以很好的过滤各种干扰。为了找到社交关系的层次关系，我们采用了自底向上的层次聚类方法，也就是以每个用户为初始的聚类中心，度量每个用户之间的亲密度，并将

司的每个部门。推荐过程中，将簇中还未加为好友的人推荐给用户。

2. **在线排序**

   在线排序是指根据用户历史行为和当前页面的上下文给推荐位的好友进行排序。好友的影响力以及与用户之间的亲密度是排序两个重要指标。这两个指标在我们的系统中作为初始的排序值。用户行为和推荐页面的上下文是我们调整推荐位好友排序值的另一因素。例如，对于展示多次但仍未引起用户注意的好友，其权值应该逐渐的衰减，新的好友推荐会排到最前面的推荐位。为了提高推荐位的点击率，我们提取用户特征，以及推荐位的特征，并根据用户的历史点击行为来预测当前推荐位的点击率，从而利用预测值来修正初始的排序，形成一个实时的反馈机制。

3. **个性化推荐**

   排序的策略保证了推荐结果的准确性，但是好友推荐往往需要结果具备多样性和新颖性。为了保证推荐结果的多样性，我们将推荐的候选集按不同的属性分成多个数据源，通过采样从不同的数据源中采集不同的好友推荐给用户。用户对不同的好友圈子的偏好可以从用户的历史行为数据中分析出来，因而可根据用户的偏好来采样，使得推荐位展示的数据符合用户偏好。

结束语

多策略融合是推荐系统提高综合性能的有效途径。推荐位的内容会越来越丰富，包括好友、公共主页、小站、音乐、日志、视频、小组等，需要建立用户模型来实现多样化推荐。用户年龄、好友数目、兴趣、活跃度等都是需要考虑的因素，根据用户的资料和行为，分析它们的统计特性，建立各个因素间依赖关系，由此构建更智能更高效的推荐引擎决策系统。

感谢张龙对本文的审校。

《架构师》月刊 架构