# 大数据信息推荐理论与关键技术

黄震华 ◎ 著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

# 大数据信息推荐理论与关键技术

黄震华 著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS
·北京·

**图书在版编目（CIP）数据**

大数据信息推荐理论与关键技术/黄震华著．—北京：科学技术文献出版社，2016.12

ISBN 978-7-5189-2115-7

Ⅰ.①大… Ⅱ.①黄… Ⅲ.①互联网络—情报检索 Ⅳ.①G354.4

中国版本图书馆 CIP 数据核字（2016）第 274337 号

**大数据信息推荐理论与关键技术**

策划编辑：赵 斌 责任编辑：赵 斌 责任校对：赵 瑗 责任出版：张志平

出 版 者 科学技术文献出版社
地 址 北京市复兴路 15 号 邮编 100038
编 务 部 （010）58882938，58882087（传真）
发 行 部 （010）58882868，58882874（传真）
邮 购 部 （010）58882873
官方网址 www.stdp.com.cn
发 行 者 科学技术文献出版社发行 全国各地新华书店经销
印 刷 者 虎彩印艺股份有限公司
版 次 2016 年 12 月第 1 版 2016 年 12 月第 1 次印刷
开 本 710×1000 1/16
字 数 178 千
印 张 11.75
书 号 ISBN 978-7-5189-2115-7
定 价 58.00 元

# 目　录

# 第1章 绪 论

## 1.1 研究背景与意义

### 1.1.1 研究背景

随着网络的快速普及和发展，每个人都是网络上信息的生产者，互联网上的内容越来越多，人们如何在海量信息中挖掘出自己想要的内容逐渐变成了一个难题。针对信息过载，已经有了大量的研究，搜索引擎的出现一定程度上缓解了这个问题。以 Google 为代表的搜索引擎通过搜集因特网上的网页信息，并利用文本信息检索等技术加以整理，最终用户能够按照关键字查询得到想要的信息。然而，搜索引擎需要用户主动、明确地提出关键字，才能检索出与关键字相关的网页，如果用户无法准确、明确地提供关键字，那么搜索引擎也无能为力。同时，搜索引擎对于所有用户的相同检索都返回一致的结果，并没有考虑到每个用户的个体差异性，缺乏个性化的结果。

信息推荐系统的出现在一定程度上能够进一步弥补这些缺陷。推荐系统是根据用户的历史偏好预测用户可能感兴趣的内容，并用一定的形式对用户做出推荐。用户在网络上的任何行为，包括浏览新闻、与好友互动、购买商品等，都可以抽象出用户的历史偏好信息。这些用户偏好被构建成用户模型，最终给用户做出个性化的推荐，每个用户获得的推荐结果与用户自身的偏好相关，从而最终满足用户的信息需求。与搜索引擎相比，推荐系统不必要求用户主动给出搜索关键字来寻找内容，而是可以主动根据用户历史偏好信息构建模型，给用户推荐出可能感兴趣的内容。

另外，长尾（the long tail）效应在互联网上显得尤为突出，信息服从长尾分布。比如，网站上售卖的商品，只有少部分被大量购买，大部分无

人问津，热销榜上的商品清单更是加剧了长尾效应，冷门物品很少被有机会展现给用户，搜索引擎获得的结果也存在这个现象，如图 1-1 所示[1]。

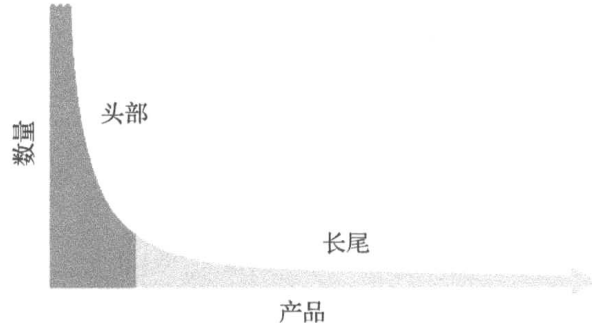


**图 1-1　长尾效应**

要想克服长尾效应，需要挖掘用户的个性化需求，而信息推荐系统返回的结果在一定程度上可以满足这个要求。信息推荐系统根据每个用户的偏好信息，计算出符合用户个性化需求的推荐结果，这样能够有效地挖掘分布在长尾处的商品。GroupLens 首次提出了协同过滤（collaborative filtering）推荐算法思想，即利用其他用户对物品的行为来预测用户对物品的偏好，为推荐系统的研究奠定了重要的基础[2]。目前，关于信息推荐系统的研究已经获得了人们的广泛关注，也取得了不少的研究成果。对于常见的用户物品推荐场景，人们提出了基于用户的协同过滤算法（user-based collaborative filtering algorithms）、基于物品的协同过滤算法（item-based collaborative filtering algorithms）及基于矩阵分解的协同过滤算法（SVD），同时，基于内容的推荐（content-based recommendation）也得到了较多的使用。这些推荐算法也在实际数据集中经历了考验，Netflix 举办的推荐比赛也大力促进了推荐系统的研究[3]。

### 1.1.2　研究意义

信息推荐系统能够挖掘出用户潜在的喜欢的内容，减少无用信息对用户的干扰，使得用户能够在互联网上快速地发现自己想要购买的商品、感兴趣的新闻、潜在的好友。并且，这些推荐结果是动态的，因为用户的兴趣会随着时间、场景的变化而发生变化，最终的推荐结果能够吻合用户的

即时偏好，给用户呈现真正感兴趣的信息内容。一个好的推荐系统不仅能够给用户推荐一些其所喜欢的信息，而且这些内容应该具有新颖性。给用户推荐的信息不应该出现多次的重复，最后推荐系统应该能够挖掘长尾信息。一个网站上的大部分商品只被少部分人所接触，一个好的推荐系统所做出的推荐结果不仅仅只是包含热门物品，而且应该多多挖掘数量巨大的冷门物品，这样才能给应用带来巨大利润和商业规模。

目前，信息推荐算法在许多商业中得到了尝试，包括电商、视频等领域。不同应用中使用的算法虽然不一样，但是大部分推荐系统的抽象框架都是相似的，都是后台通过分析用户行为构建用户模型，然后使用推荐算法计算出推荐结果，最终通过各种具体的交互页面呈现给用户。亚马逊是一个在线购物网站（图1-2），很早就运用了推荐系统，即通过分析用户行为，给用户推荐商品。


图1-2 亚马逊推荐

在首页中根据每个用户的点击、浏览、购买、评价等历史行为，有两种推荐：一种是“您浏览过商品相关的推荐”，主要根据那些用户浏览过物品的相关性做出相似物品推荐；另一种是“根据您的浏览历史记录推荐商品”，则是个性化的推荐。推荐系统在用户购物中的使用，使其销售额提升了30%[4]。

Netflix是一个视频网站，为了给用户提供个性化的推荐结果，其举办了奖金高达100万美元的推荐系统比赛，并开放了他们网站的真实数据集，即40多万用户对2万多部电影的评分记录。经过参赛者的共同努力，

最终的推荐效果提升了超过 10%。同时，这次比赛在推荐系统研究领域产生了不小的影响，比赛中的公开数据集也被广泛应用在学术研究中[5]。

Facebook 搭建了 Apache Giraph 推荐平台，能够处理、分析超过 10 亿用户及数百万的物品数据，并给用户做出精确的推荐。同时，Facebook 中还利用 FOFs 推荐系统，综合用户之间的社交关系及用户特征构建模型给用户推荐可能感兴趣的朋友[6]。

信息推荐系统在音乐推荐中也得到了广泛的研究和应用，比如个性化的豆瓣电台（图 1-3），给用户提供了对歌曲进行“喜欢这首歌”、“下一首”、“不再播放”等行为操作，最终通过这些行为给用户推荐一些歌曲。


**图 1-3　豆瓣 FM 推荐**

## 1.2　国内外研究现状及分析

### 1.2.1　国外研究现状及分析

现阶段协同过滤推荐、基于内容推荐及基于矩阵分解推荐等算法得到了广泛的使用。协同过滤推荐算法仅仅利用用户对物品的评分，抽象出用户向量、物品向量，进而计算出用户之间的相似度或者物品之间的相似度，然后根据这些相似用户、物品预测用户对物品的评分[7]。矩阵分解目前被广泛研究。矩阵分解是将用户物品评分矩阵 $[R_{u,i}]_{N\times M}$ 分解为低维度的 2 个矩阵（$U$ 和 $V$），即 $R\approx U^TV$，$U\in R_{d\times m}$，$V\in R_{d\times n}$，这里 $d$ 是一个

远远小于 $m$、$n$ 的维度，通过分解后的低维度用户向量和物品向量来计算该用户对该物品的偏好[8]。协同过滤和矩阵分解都只是利用了用户物品评分数据，没有考虑用户之间的其他关系，因此不能够进一步有效地提升准确度。同时，传统的推荐方法也存在着其他一些问题，比如在计算两个用户之间的相似度时，需要考虑这两个用户对所有物品的评分记录。如果两个用户共同评分的物品比较少，甚至没有，这种情况是很常见的，比如在电子商务或者在线电影评分场景中，商品或者电影的数量可能有上百万之多，而每个用户评分过的商品、电影可能不到 10 个。这时随机选择两个用户计算相似度时，他们很可能没有共同评分过的物品，这样计算出来的用户之间相似度不是很准确，对后面依赖用户相似度的计算推荐结果也会产生负面影响。同时，传统的推荐中没有考虑用户之间的关系，如果一个恶意用户登录注册网站后，对大部分商品进行了大量的评分评价行为，那么这个恶意用户的评分最终也会被推荐模型所考虑，而这正是恶意用户想要的对推荐网站的恶意攻击行为。

进一步，研究人员发现社交网络在推荐系统中扮演着重要的角色，除了仅仅考虑用户的物品评分信息，还可以考虑用户之间的社交关系来进一步提高推荐的准确度。社交关系是指用户之间的好友关系或者关注关系，网络中的社交关系一定程度上体现着用户在实际生活中的交际情况。可以定义社交网络 $G$（$U$，$E$，$W$），其中，$U$ 是顶点集合，表示社交网络中的所有用户，$E$ 是边集合，如果 $u_a$ 和 $u_b$ 两个用户之间有社交关系，则存在边 $e_{a,b}$连接这两个用户，$w_{i,j}$即是其值的大小，表示这个社交关系的权重。

Barbieri N 等人[9]提出假设：用户间的社交关系可以分为 topical 和 social 两类，同时每个用户也会属于一个或多个社交圈子，而每个社交圈子都有一定的概率属于 topical 和 social，然后利用概率模型预测用户之间的潜在社交关系来给用户做出推荐，并给出推荐解释。

虽然在参考文献[9]中，作者将用户社交关系分为 topical 和 social 两类，并取得了较好的效果，但是用户行为的复杂性很难用两种或者几种社交关系来度量，用户的社交关系也很难准确地判定。而用户之间的信任度能够有效反映用户之间的关系，这对计算个性化的推荐结果很有帮助。Jamali M 等人[10]指出用户的行为被所信任的人影响，因此表示用户信息的向量为：

$$\hat{U}_u = \frac{\sum_{v \in N_u} T_{u,v} U_v}{\sum_{v \in N_u} T_{u,v}}$$

式中：$N_u$ 表示与用户 $u$ 有信任度关系的用户集合，$T_{u,v}$ 表示用户 $u$ 和 $v$ 间的信任度值，$U_v$ 表示用户 $v$ 的社交朋友集合。这样计算出的用户向量 $\hat{U}_u$ 由它的邻居及信任度传播所决定，然后利用矩阵分解来预测评分。

### 1.2.2 国内研究现状及分析

Shen Y 等人[11]从用户的社交关系中抽取出表示用户社交关系的向量，同时结合协同过滤进行矩阵分解，最后在爬取的豆瓣数据集和 flixster 数据集上做了实验，来验证矩阵分解后的维度因素对 MAE（平均错误率）和 RMSE（均方根误差）的影响。实验结果证实：在推荐系统中考虑用户之间的社交关系能够有效提升推荐的效果。虽然用户的显示社交关系能够有效地改善推荐效果，但是很多时候数据集中并没有显示的社交关系。其实隐式的社交关系也能够对提高推荐系统的准确率做出一定的贡献。Ma H 等人[12]指出：如果两个用户相似度较高，那么代表这两个用户的向量应该相近，利用用户之间的相似度来模拟用户好友关系，并在矩阵分解的目标函数中根据用户之间的相似度来约束代表这两个用户的向量的差值。实验表明：利用这种挖掘出的隐式的社交关系做出的推荐效果，比利用显式的社交关系的推荐效果只差一点点。但是，用户的社交关系背后其实包含了多种因素，很多时候用户之间的共同爱好并不是促成好友的唯一因素，比如一个用户关注另一个用户可能是因为他们是同事关系，或者是被关注用户在某领域具有权威性，社交关系背后含义的不同对推荐结果产生着巨大的影响。邹本友等人[13]提出了一种结合用户信任度及张量分解的推荐算法，能够针对新加入推荐系统的用户物品进行增量更新，并在不同稀疏度的数据集上验证了模型的准确度情况，这种方法能够提高动态变化的社交网络推荐效率。参考文献[14]指出：在用户的社交网络中，不仅正向的关系、信任度信息可以用于计算推荐结果，用户之间的负面关系，比如不信任度也能够提高推荐结果的准确率。

## 1.3 信息推荐目前存在的主要问题

随着近年来对信息推荐系统的深入研究，许多研究难点与挑战得到了大家的广泛关注，主要存在如下5个问题。

### 1.3.1 信任度传播问题

研究表明，人们往往更容易相信来自他所信任的朋友的推荐，而不是一个普通陌生人的推荐，尽管这个陌生人与他的相似度较高。网络中用户之间的信任度是现实中用户联系的反映，体现着用户之间相互信任的程度。如果与一个用户的信任度较高的朋友对一个物品给予了高度评价，那么这个用户很有可能会接受这个物品的推荐。不仅用户的推荐结果会受到他所信任的朋友的影响，而且也会受到他所信任的朋友的朋友的影响，这就是信任度传播。

信任度网络中，虽然每个用户只与少数几个近邻用户有着直接相连的关系，但是近邻用户也会和其他用户有着直接关系，那么这个用户其实与近邻的近邻用户也有着一定程度的间接关系。基于信任度的社交网络中的信任度传播，使得在计算推荐结果时能够有效地参考信任用户的偏好行为，从而使得推荐结果更加准确，推荐的覆盖率更高。但是信任度的传播也带来了问题：信任度传播距离越远，到达的用户越多，虽然源用户能够获得更多可用的信息来做出个性化的推荐，但是噪声也越大，推荐结果的精确性受到负面影响；传播距离越近，到达的用户越少，则不能为源用户提供足够多的可利用信息，不能进一步有效提高准确率与覆盖率。

多样性、新颖性等指标很少在传统的推荐算法中被考虑，给用户推荐一个意料之外的可能喜欢的物品远远比向用户推荐一个热门、流行物品更能让用户感到惊喜。而且提高推荐结果的多样性能够挖掘物品的长尾分布，使得推荐结果对于用户而言有更好的新颖性，这能更好地提升推荐系统的性能。

### 1.3.2 执行效率问题

目前的推荐算法大都需要经过大量复杂的计算才能得到推荐结果，而

社交网络推荐算法为了参考其他用户的偏好信息，需要遍历社交网络中的用户节点。然而在实际工业应用中，由于用户的社交网络是动态变化的，在海量数据面前，传统的推荐算法显得无能为力、效率太低，很难在生产中得到普遍应用，因此，设计分布式的推荐算法显得极为迫切。分布式计算框架可以将大的任务分成若干个子任务并行处理，之后将得到的各部分结果进行汇总，得到最终的计算结果。本书对提出的推荐算法进行了并行设计，基于 Hadoop 平台，对社交网络划分出用户社区，并优化了每个节点的输入，最终在保证计算结果准确度的同时极大地提高了算法的执行效率。

### 1.3.3 实时性问题

企业知名度的提高及其业务量的增加，使得电子商务网站上的用户资料和经营的产品信息频繁发生变化，例如，eBay 网每天有近 4000 万的浏览量及数百万的新刊登物品。这样，一旦用户偏好发生改变，现有的推荐方法需要花费巨大的时间开销来重新进行兴趣建模并产生推荐结果，这将严重影响信息推荐的实时性。

### 1.3.4 推荐质量问题

在现实应用中，用户受到不同因素的影响，其偏好经常会发生变化。这种偏好的演变过程很大程度上蕴含着该用户消费的隐性知识，因此，分析和挖掘其规律能够很好地预测用户将来的偏好趋向。然而，现有的信息推荐方法通常只对用户的当前偏好进行建模，而不关心偏好的演变过程，这在很大程度上将影响推荐的质量和效果。

### 1.3.5 鲁棒性问题

近年来，信息推荐系统的研究核心是推荐方法，而较少涉及信息推荐系统的鲁棒性研究。然而随着 Web 2.0 技术的成熟，网络开放特性使得网站经常受到恶意用户的攻击，以及用户流量压力导致的软件模块异常，致使信息推荐系统不稳定和不可靠，并最终导致系统瘫痪。

## 1.4 本书结构与研究内容

本书的组织结构如下：

第1章为绪论部分。

第2章在分析传统3类推荐算法存在问题的基础上，介绍和分析了语义推荐算法的研究现状和进展，主要包括基于语义的内容推荐算法、基于语义的协同过滤算法、基于语义的混合推荐算法及基于语义的社会推荐算法，并讨论了语义推荐算法所面临的主要挑战及今后的研究发展方向。

第3章主要分析和研究了基于排序学习推荐算法的现状、关键技术和进展，并讨论了今后的发展方向。3.2节给出了基于排序学习推荐算法的框架；3.3节给出了基于排序学习推荐算法的关键技术，包括点级排序学习、对级排序学习及列表级排序学习等相关技术；3.4节给出了排序学习的效用评价准则，包括准确率、召回率、MAP（平均准确率）、NDCG（Normalized Discounted Cumulative Gain，用来衡量排序质量的指标）等；3.5节给出了基于排序学习推荐算法的应用进展，包括微博服务、多媒体应用、标签推荐、个性化新闻推送、购物推荐及移动推荐；3.6节分析了基于排序学习推荐算法的研究趋势。

第4章主要针对云环境下的top-n推荐算法进行了深入研究，提出了适合top-n推荐的多层分布式存储架构（Multilayer Distributed Storage Architecture，MDSA），并从降低网络传输代价出发，设计了基于MDSA架构的数据编码模式，进而利用map/reduce分布式编程模型快速实现top-n推荐。此外，为了满足实际需求，给出了3种top-n推荐的应用扩展。

第5章主要解决轮廓推荐预处理阶段数据传输量的问题。基于全空间轮廓集合与子空间轮廓集合之间的语义关系，提出一种三阶段传输查询所需数据的有效方法TPAOSS。在TPAOSS方法的第一阶段，只传送全空间轮廓对象；在第二阶段，接收超点计算机回送的种子轮廓对象的位置值；而在第三阶段，基于Bloom Filter技术来传送种子轮廓对象在子空间上的重复对象位置值及重复频率。

第6章分析了现有工作存在的主要性能缺陷，并给出一种在SPA分

布式网络中进行轮廓推荐的有效方法——EMSRDN。EMSRDN 算法不以底层细粒度的数据为输入参数，而采用预存储 $w$ 个轮廓快照来高效处理系统中的 $u$ 个轮廓推荐指令，并且利用 map/reduce 分布式计算模型，通过初始轮廓快照集启发式构造与基于遗传算法的轮廓快照集深度优化这两个阶段来快速产生最优的 $w$ 个轮廓快照。

第 7 章放松了传统轮廓推荐的限制，并给出一类新的轮廓推荐应用，即 k-quasi 轮廓推荐，来丰富传统的轮廓推荐结果集。7.2 节给出即 k-quasi 轮廓推荐形式化语义；7.3 节给出了一种基于正规格索引的新颖算法——EARG，来有效处理任意维空间 k-quasi 轮廓推荐；7.4 节提出了维空间推荐树的概念，并利用树路径上各节点所对应 k-quasi 轮廓集合间的关系来缩减多个 k-quasi 轮廓推荐的总响应时间；7.5 节基于其他不同的实际应用，讨论了两个 k-quasi 轮廓推荐的变种，即全局约束 k-quasi 轮廓推荐和局部约束 k-quasi 轮廓推荐，来丰富传统轮廓推荐的语义。

第 8 章将基于密度的聚类技术引进到实际的轮廓推荐应用中，提出轮廓类操作符的概念。另一方面，为了在大数据环境下有效实施轮廓类操作符，该章给出一种有效聚类轮廓对象集的方法——EAPSC。EAPSC 算法在处理轮廓推荐的过程中，将各轮廓对象的位置信息组织成一棵新颖的 SLT 索引树，并且在聚类轮廓对象集的过程中，使用 SLT 索引树多个有效的剪枝性质来快速产生所有的轮廓类。

第 9 章提出基于正规格结构的 K-均值类推荐算法 KMCRG。9.2 节介绍了 K-均值类推荐算法的核心思想；9.3 节给出了 KMCRG 算法的工作原理。KMCRG 算法通过将多维数据集划分为若干个组，每个组中的对象均落入同一个单元格中，使得我们能够以单元格为处理单位来取代传统以单个多维对象为处理单位的做法，从而有效降低 K-均值类推荐的时间开销。为了不损失聚类的精度，KMCRG 算法分为两个阶段来返回最终的 $K$ 个聚类：在第一阶段中，利用信息熵理论来有效进行初始聚类中心的选择，而在第二阶段中，使用单元格加权迭代的策略来有效返回最后的结果。

# 第2章　大数据语义推荐算法

## 2.1　引言

近年来，随着物联网、云计算和社会网络等技术的迅猛发展，网络空间中所蕴含的信息量将呈指数级增长[15]。例如，Facebook 每月上传的照片超过 10 亿张，每天生成 300TB 以上的日志数据；淘宝网站每天有数千万笔交易，单日数据产生量超过 50TB；YouTube 线上有数千万部电影，每天要处理上千万个视频片段；AOL Music 在线音乐网站的音乐库包含有 250 万首歌曲和数千首音乐电视，每天独立用户访问量达到 2500 万。不难发现，信息过载呈爆发趋势，其结果导致了终端用户无法准确和高效地获取自己感兴趣的物品[16,17]。

目前，推荐系统是解决信息超载问题最有效的工具之一[18]。推荐系统的概念是 AT&T 贝尔研究院的 Paul R 博士在 1997 年提出的，即通过建立用户与物品之间的二元关系，利用用户的历史记录或物品之间的相似性关系，挖掘每个用户潜在感兴趣的物品[19,20]。不难看出，推荐系统由 3 个基本要素组成，即用户、产品和推荐算法，而推荐算法是推荐系统的核心部分，它决定着推荐系统性能的优劣[21]。传统的推荐算法可以归纳为 3 个大类，即基于内容的推荐算法、协同过滤推荐算法及混合推荐算法[22,23]。

2006 年，Loizou A 博士在意大利特兰托市召开的推荐系统研讨会（ECAI 2006 Recommender Systems Workshop）上指出：传统的推荐算法由于没有考虑应用场景的上下文语义，在实际应用中，这些算法在实时性、鲁棒性和推荐质量等方面存在严重的不足。其还提出了语义推荐算法的概念，其核心思想是将语义知识融合到推荐过程中，来克服传统推荐算法的不足[24]。随后，许多学者开始将语义技术集成进各类传统推荐算法中，