大数据挖掘技术及其应用
Big Data Mining Technology and Applications
第9章
推荐系统

内容概要

教学大纲要求

- 推荐系统的模型
- 基于内容的推荐系统
- 基于协同过滤的推荐系统

教学基本要求：掌握推荐系统基本概念，基于内容的推荐系统和基于协同过滤的推荐系统。

推荐系统导言

推荐系统（Recommendation System，简称RS）技术，它根据用户的兴趣、行为、情景等信息，把用户最可能感兴趣的内容主动推送给用户。近年来，推荐系统技术得到了长足的发展，不但成为学术研究的热点之一，而且在电子商务、在线广告、社交网络等重要的互联网应用中大显身手。

推荐系统导言

人们在日常工作和决策时经常采用找朋友聊聊、从可信的第三方获取信息、在互联网上咨询、凭直觉或索性随大流等方法获得建议。

上述方法带来的决策并不那么有效，大多数情况下，花费了大量的时间和金钱，结果总是让人半信半疑，例如：推销员大献殷勤的建议并不那么有用；凭感觉跟着邻居或好友投资，却没有真正给我们带来收益；无休止地花费时间在互联网上会导致困惑，却不能做出迅速而正确的决定。

推荐系统导言

随着Web技术的发展，每天都有大量的图片、博客、视频发布到网上。一方面，内容的创建和分享变得越来越容易，另一方面，互联网信息的爆炸式增长和种类的纷繁复杂使得人们找到他们需要的信息、作出最恰当的选择是非常困难的。如何解决信息过载问题？

搜索引擎
推荐系统

推荐系统导言

推荐系统和搜索引擎的异同点：

- 相同点：
  - 都是一种帮助用户快速发现有用信息的工具
- 不同点：
  - 搜索引擎需要用户主动提供准确的关键词来寻找信息
  - 推荐系统不需要用户提供明确的需求，而是通过分析用户的历史行为给用户的兴趣建模
- 从某种意义上说，推荐系统和搜索引擎对于用户来说是两个互补的工具
  - 搜索引擎满足了用户**有明确目的时的主动查找**需求
  - 推荐系统能够在用户没有明确目的的时候**帮助**他们**发现感兴趣的新内容**

推荐系统导言

发展历史

相比于其他经典的信息系统的工具和技术，如数据库和搜索引擎，推荐系统的研究是相对较新的。在20世纪90年代中期，推荐系统成为一个独立的研究领域。回顾推荐系统发展过程，到目前为止可分为四个阶段：

1.探索性阶段

本阶段主要以早期的协同过滤系统为代表。如Tapestry系统，GroupLens系统。

这一阶段的标志性实践是1996年3月在伯克利举办的协同过滤专题研讨会

推荐系统导言

发展历史

2.商业化阶段

信息大爆炸的互联网环境下，人们对精准有效信息的渴求催生了推荐系统的出现，因此，推荐系统的商业化几乎刻不容缓。MIT的Pattie Maes研究组于1995年创立了Agents公司（后来更名为萤火虫网络，Firefly Networks）。美国明尼苏达州的GroupLens研究组于1996年创立了Net Perceptions。

在商业化推进的过程中，推荐系统的商业化应用遇到了实验室里未曾面临的真实挑战：必须在不降低现有Web站点速度的情况下证明能够提供有价值的推荐，这些系统必须能够在大大超越实验室规模的情况下运行（处理上百万的用户和物品以及每秒成百上千的交易）。

推荐系统导言

发展历史

3.大爆发阶段

2000至2005年间，一方面，互联网泡沫逐渐破灭，另一方面，推荐系统被整合到更全面的商业产品线的主流公司，许多专用的推荐系统公司逐渐消亡了。然而，推荐系统作为一门技术仍然存在，并广泛应用在电子商务、大规模零售业和各种知识管理应用中。与此同时，随着各个学科研究人员的参与及方法的引入，推荐系统研究得到迅猛发展。来自人工智能、信息检索、数据挖掘、安全与隐私以及商业与营销等各个领域的研究，都为推荐系统提供了新的分析和方法。由于可以获取到海量数据，算法研究方面取得了很大进步，在2006年更是被悬赏100万美元将预测精确度提高到10%的Netflix大奖推上高峰。

推荐系统导言

发展历史

4.再前进阶段

由于推荐系统实际应用效果显著，近年来国际学术界与其相关的研究极为活跃。2006年，MyStrands组织了Recommenders06大会，这是一个介绍推荐系统现状和未来的暑期班。推荐系统研究的顶级会议是美国计算机学会（ACM）每年举办的RecSys年会，该会议自2007年以来每年举行一次，成为全球关于推荐系统研究的最重要的交流渠道和把脉其最新进展的重要窗口。上述事件揭示了人们对于基于上下文的推荐越来越感兴趣，乐于改进研究方向使其立足于理解人们如何与机构或企业互动。

推荐系统导言

评价标准

实时性：在很多网站中，因为物品（新闻、微博等）具有很强的时效性，所以需要在物品还具有时效性时就将它们推荐给用户。比如，给用户推荐昨天过时的新闻显然不如给用户推荐今天刚刚发生的新闻。因此，在这些网站中，推荐系统的实时性就显得至关重要。

健壮性：具有经济效益的算法系统常常会受人攻击，以搜索引擎为例，如果某个商品称为热门搜索词的第一个搜索结果，将会带来极大的商业利益，因此，搜索引擎的作弊和反作弊斗争异常激烈。目前，推荐系统也遇到了同样的作弊问题，而健壮性（即robust，鲁棒性）指标衡量了一个推荐系统抗击作弊的能力。

9.1 一个推举系统的模型

两个很好的推荐系统样例:

- 基于对用户兴趣的预测结果，为在线报纸的读者提供新闻报道;
- 基于顾客过去的购物和/或商品搜索历史，为在线零售商的顾客推荐他们可能想要买的

推荐系统可以分成两大类。

- 基于内容的系统(Content-based System)：这类系统主要考察的是推荐项的性质。项之间的相似度通过计算它们的属性之间的相似度来确定。例如，如果一个Netflix的用户观看了多部西部牛仔片，那么系统就会将数据库中属于“西部牛仔”类的电影推荐给该用户。
- 协同过滤系统(Collaborative Filtering System)：这类系统通过计算用户或/和项之间的相似度来推荐项。与某用户相似的用户所喜欢的项会推荐给该用户。

9.1 一个推举系统的模型

9.1.1 效用矩阵

在推荐系统应用当中，存在两类元素，一类称为用户(user)，另一类称为项(item)。

数据本身会表示成一个效用矩阵(utility matrix)，该矩阵中每个用户-项对所对应的元素值代表的是当前用户对当前项的喜好程度。

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

9.1 一个推举系统的模型

9.1.1 效用矩阵

推荐系统的目的是预测效用矩阵的空白元素。比如，用户A是否喜欢SW2？从效用矩阵中无法获得证明，但我们可以在涉及效用矩阵的时候考虑电影的属性，比如制片人、导演、演员等。我们可能会发现SW1和SW2的相似性，从而由于A不喜欢SW1，那么他可能也不喜欢SW2。

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

9.1 一个推举系统的模型

9.1.2 长尾现象

物理世界和在线世界的差别被称为长尾现象，物理机构只列出图中竖线左部的最流行项，而相应的在线机构则会提供全范围的项，即不仅包括流行项也包括尾部项。

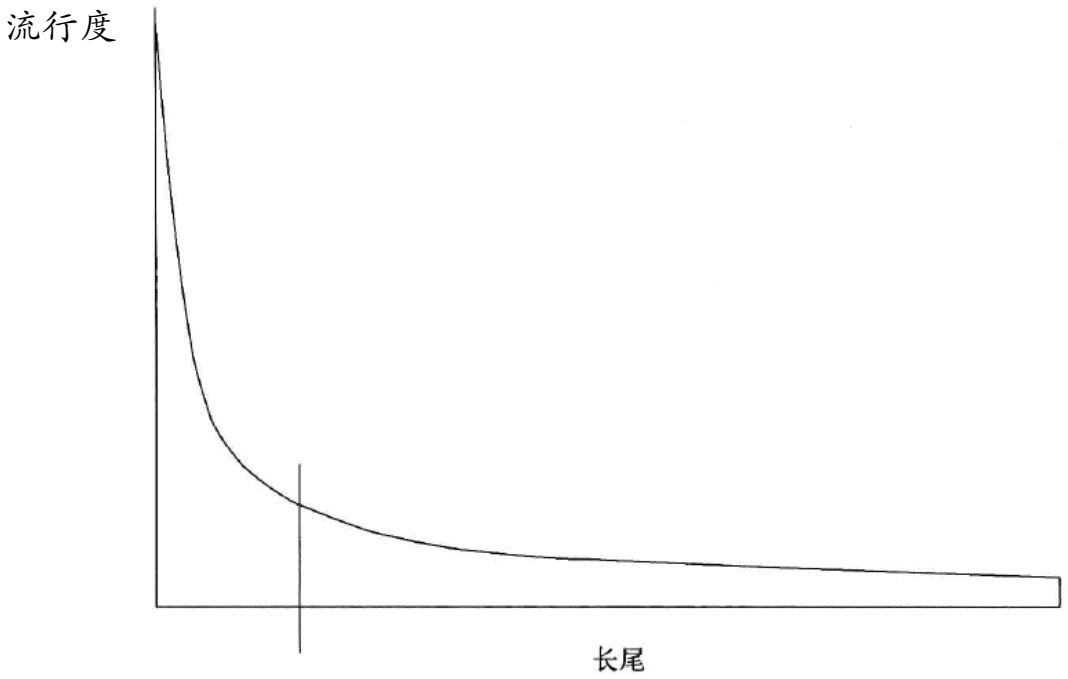


图9-2　长尾：物理机构只能提供流行项，而在线机构可以提供所有项

9.1 一个推举系统的模型

9.1.2 长尾现象

长尾现象要求在线机构必须对每个用户进行推荐。将所有项推荐给用户是不太可能的，这里和物理机构的情况类似，即期望用户听说过他们喜欢的所有项也是不可能的。

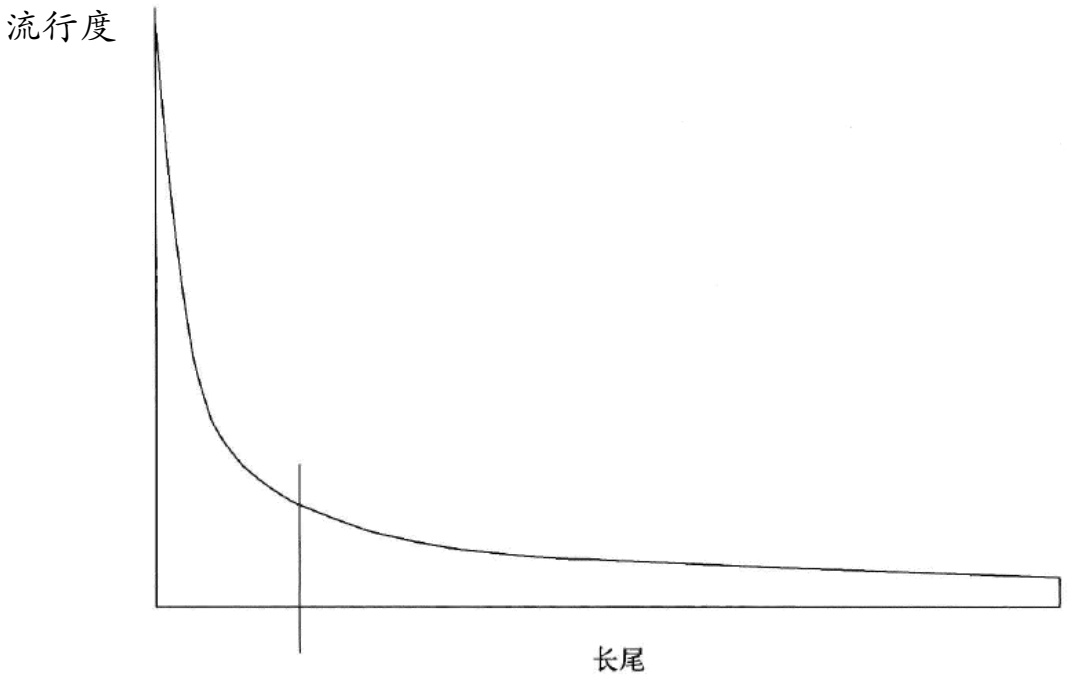


图9-2　长尾：物理机构只能提供流行项，而在线机构可以提供所有项

9.1 一个推举系统的模型

9.1.3 推荐系统的应用

- 产品推荐 或许最重要的推荐系统应用于在线零售商。
- 电影推荐 Netflix会为其用户推荐他们可能喜欢的电影。
- 新闻报道 推荐新闻服务机构已经试图基于读者过去所阅读的文章来识别读者的兴趣。

9.1.3 4 效用矩阵的填充

- 我们可以邀请用户对项评级 （有限）
- 我们可以根据用户的行为来推理 （更加实际）

回归模型，字典模型，low rank 等等

9.2 基于内容的推荐

推荐系统主要有两类基本的架构

- 基于内容的系统集中关注项的属性。项之间的相似度通过计算它们的属性之间的相似度来确定。
- 协同过滤系统集中关注用户和项之间的关系。

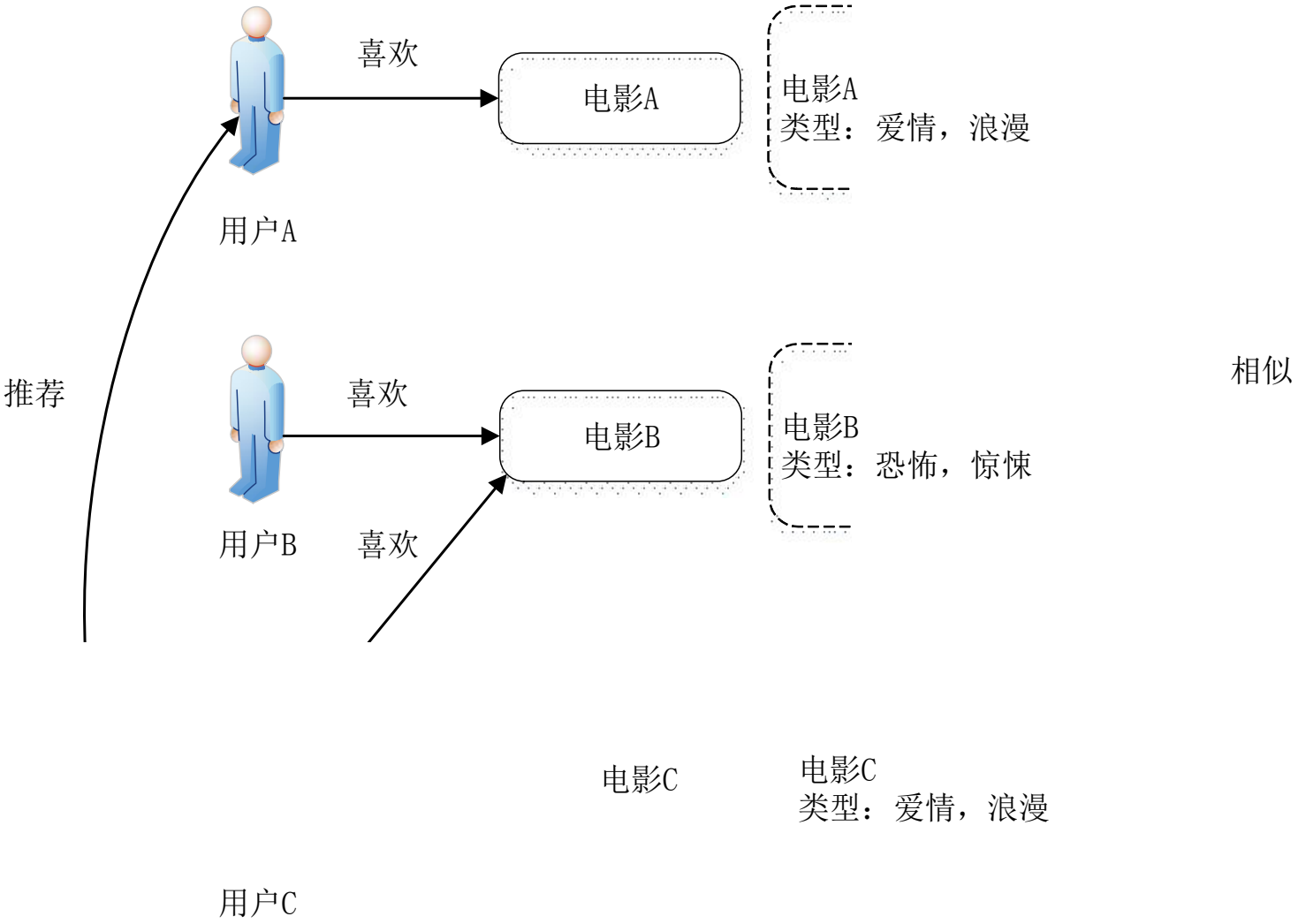
基于内容的推荐系统是在推荐系统出现之初应用最为广泛的推荐机制；
它的核心思想是挖掘用户曾经喜欢的物品，从而尝试去推荐类似的物品使用户满意。
具体来说，基于内容的推荐系统通过分析一系列用户之前已评分的文档和（或）描述，从而基于用户已评分物品的特征建立用户个人信息。

9.2 基于内容的推荐

在一个基于内容的系统中，我们为每个项建立一个模型，用于代表该项的重要特性的一条或多条记录。例如，构建电影的如下特征：

1. 演员集合。一些用户会偏向他们喜欢的演员所出演的电影。
2. 导演。一些用户会偏向某些导演的电影作品。
3. 年份。一些用户偏好最新电影，一些用户偏好老电影。
4. 流派。一些用户只喜欢喜剧电影，一些用户喜欢科幻电影。

9.2 基于内容的推荐


喜欢
电影A
电影A
类型：爱情，浪漫
用户A
推荐
喜欢
电影B
电影B
类型：恐怖，惊悚
相似
用户B
喜欢
电影C
电影C
类型：爱情，浪漫
用户C

9.2 基于内容的推荐

基于内容推荐存在的问题

需要对物品进行分析和建模，推荐的质量依赖于对物品模型的完整和全面程度。在现在的应用中我们可以观察到关键词和标签（Tag）被认为是描述物品元数据的一种简单有效的方法。

物品相似度的分析仅仅依赖于物品本身的特征，这里没有考虑人对物品的态度。

因为需要基于用户以往的喜好历史做出推荐，所以对于新用户有“冷启动”的问题。（这里的冷启动是指用户冷启动，当新用户到来时，没有他的行为数据，所以也无法根据他的历史行为预测其兴趣，从而无法借此给他做个性化推荐。）

9.2 基于内容的推荐

9.2.1 项模型

一个基于内容的系统中，我们必须为每个项建立一个模型(profile)，即用于代表该项的重要特性的一条或多条记录。

以电影为例

- 电影中的演员集合 用户会偏向他们喜欢的演员
- 导演 一些用户会偏向某些导演的电影作品
- 电影的流派(genre)或常规类型。一些用户只喜欢喜剧电影，而有些用户则喜欢剧情片 或爱情片。

互联网电影数据库(Internet Movie Database, IMDB)

9.2.2 文档的特征发现

TF.IDF (绪论)

TF：词条在文档中的出现频率
IDF：文档总数与包含目标词条的文档数的比值

为了解决简单布尔方法的缺陷，一种实际有用的做法是从文档中找出能够刻画主题的**关键词**。

例如，有关足球（football）的文章当中往往会出现类似“ball”（球）、“forward”（前锋）、“midfield”（中场）、“back”（后卫）、“Corner”（角球）之类的词语。如果将文档分到确实是关于足球的主题类中，上述词语在文档中可能会十分频繁。然而，我们不能纯粹地从词语在文档中出现的频繁程度来断定该词语刻画了文档的主题类别。例如，在英文文档中，出现最频繁的大部分词语都是类似“the”或者“and”的常见词（这些词通常都用于辅助表达但本身不携带任何意义，又称为停用词）。因此，英文文档在进行分类之前往往会先将上述停用词去掉。

9.2 基于内容的推荐

9.2.3 基于Tag的项特征获取

考虑图像数据库上的特征获取方法

- 邀请用户采用词语或短语对图像进行标记，那么就可以从这些标记中获得有关图像特征的信息。
- 基于计算机游戏的标记过程（比如RPG游戏里的地图标记）
- 基于人工智能的标记

9.2.4 项的模型表示

在基于内容的推荐中，我们的最终目标是构建由特征-值对构成的项模型，并基于效用矩阵中的每一行构建反映用户偏好的用户模型。

$$0\quad 1\quad 1\quad 0\quad 1\quad 1\quad 0\quad 1\quad 3\alpha$$

$$1\quad 1\quad 0\quad 1\quad 0\quad 1\quad 1\quad 0\quad 4\alpha$$

9.2 基于内容的推荐

9.2.5 用户模型

我们不仅要为项建立向量表示，也需要将用户的偏好表示成同一空间下的向量。

我们拥有将项和用户关联起来的效用矩阵。

我们还记得，效用矩阵中的每个非空元素可以代表用户购买过该项(表示为1)或类似关系，也可以是表示用户对项的评分或喜好程度的一个任意数字。

其他知识点

9.2.6 基于内容的项推荐

9.2.7分类算法

9.3 协同过滤

协同过滤是目前研究最多也是应用最成熟的个性化推荐技术，是与基于内容的推荐完全不同的一种推荐方法。

基于内容方法是使用被用户评过分的物品内容，协同过滤方法还取决于被其他用户评分过的物品内容。通过分析用户评价信息（评分）把有相似需求或品味的用户联系起来，用户之间共享对物品的观点和评价，这样就可以更好地做出选择。

例如，当你在网上买衣服时，基于协同过滤的推荐系统会根据你的历史购买记录或是浏览记录，分析出你的穿衣品位，并找到与你品味相似的一些用户，将他们浏览和购买的衣服推荐给你。

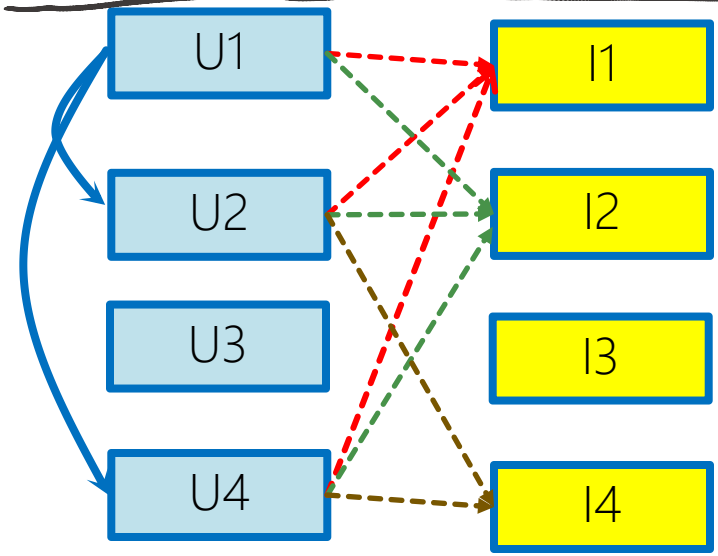
Typestry是最早提出来的协同过滤推荐系统，用于过滤电子邮件，推荐电子新闻,由于其要求用户手工输入查询条件，不牵涉到用户间的相似性计算，严格来讲，它只是一个信息检索系统，只是对检索结果根据其它用户的反馈进行筛选，其它的协同过滤推荐系统有GroupLens/NetPerceptions,Ringo/Firefly等。

9.3 协同过滤

协同过滤- 使用行为数据，利用集体智慧推荐

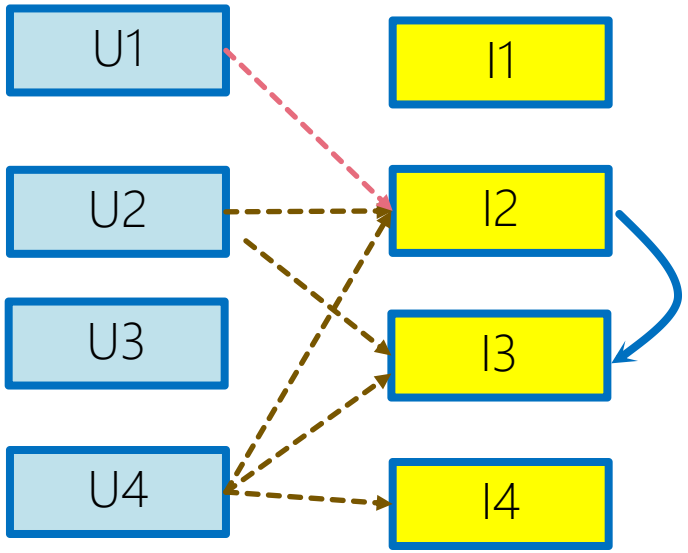
基于用户的协同过滤 U2U2I

和你兴趣相投的人也喜欢XXX



基于物品的协同过滤 U2I2I

喜欢这个物品的人也喜欢XXX



9.3 协同过滤

协同过滤 实例

步骤1：搜索最相似用户

步骤2：计算用户和新的项的相似度

	I1	I2	I3	I4	I5	I6	I7
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

相似度计算：

Jaccard相似度：缺点没有考虑评分

Sim(A,B)=1/5

Sim(A,C)=2/4

9.3 协同过滤

协同过滤 实例

步骤1：搜索最相似用户

步骤2：计算U和新的item的相似度

	I1	I2	I3	I4	I5	I6	I7
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

相似度计算：

Jaccard相似度：缺点没有考虑评分

Sim(A,B)=1/5

Sim(A,C)=2/4

余弦相似度：缺点缺失值为0

Sim(A,B)=0.38

Sim(A,C)=0.32

9.3 协同过滤

协同过滤 实例

步骤1：搜索最相似用户

步骤2：计算U和新的item的相似度

	I1	I2	I3	I4	I5	I6	I7
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

相似度计算：

Jaccard相似度：缺点没有考虑评分

Sim(A,B)=1/5

Sim(A,C)=2/4

余弦相似度：缺点缺失值为0

Sim(A,B)=0.38

Sim(A,C)=0.32

皮尔逊相关系数：缺失值为平均值

Sim(A,B)= 0.09

Sim(A,C)=-0.56

9.3 协同过滤

协同过滤 实例

步骤1：搜索最相似用户

步骤2：计算U和新的item的相似度

	I1	I2	I3	I4	I5	I6	I7
A	4			5	1		
B	5	5	4			3	
C				2	4	5	
D		3					3

方案1：直接求平均值

Sore（a，I5）=（3+5）/2

9.3 协同过滤

协同过滤 实例

步骤1：搜索最相似用户

步骤2：计算U和新的item的相似度

	I1	I2	I3	I4	I5	I6	I7
A	4			5	1		
B	5	5	4			3	
C				2	4	5	
D		3					3

方案1：直接求平均值

Sore（a，I6）=（3+5）/2

方案2：加权平均

Sore（a，I6）=（w_{ab}x3+w_{ac}x5)/(w_{ab}+w_{ac})

9.3 协同过滤

9.3.2 相似度对偶性

问题描述：

（1）我们可以使用与用户相关的信息来推荐项。也就是说，给定用户，我们可以找到最相似的一些用户

（2）典型用户的行为和项的行为有一点不同，这与相似度计算有关。

权衡：

（1）如果寻找相似用户，那么我们只需要对用户U做一次。基于相似用户集合我们可以估计效 用矩阵中U那行的所有空白元素。

（2）发现同一流派的项要比发现只喜欢单个流派项 的用户要容易得多，因此，项之间的相似度常常可以提供更可靠的信息。

注意： 协同过滤是一个推荐算法的**基本架构/框架**，可以有多种技术实现方式

9.3 协同过滤

9.3.3 用户聚类和项聚类

面对日益增多的用户，数据量的急剧增加，算法的扩展性问题(即适应系统规模不断扩大的问题)成为制约推荐系统实施的重要因素。聚类技术是比较公认的有效减少项目数的办法：

（1）k-means 聚类算法

（2）Expectation-Maximization 算法（EM）：该算法首先估计用户或项目属于某一类的概率，然后按照概率值对用户或项目进行聚类。

（3）Gibbs Sampling 方法：　该方法的原理与 EM 算法有些类似，都是首先估算概率值，所不同的是，Gibbs Sampling 方法需要估计三个概率参数：Pk、Pi 和Pki，然后再利用这三个参数对用户或项目进行聚类。

（4）模糊聚类：　该方法通过设定固定的阀值，并以此来确定对象的相似类别，其与聚类的区别仅仅在于不需要预先给定聚类的数目。。

9.3 协同过滤

补充 协同过滤的优缺点

优点：

（1）对于很难自动分析非结构化数据的项目（如艺术品、音乐等），该算法能够有效过滤。

（2）克服了基于内容推荐算法依赖用户和项目特征的缺点，不需要预处理项目和用户的特征。

（3）这种算法可以推荐出用户的潜在需求，从而能推荐新产品售卖。

9.3 协同过滤

补充 协同过滤的优缺点

缺点：

（1）个人隐私问题：协同过滤推荐系统需要用户大量的行为数据和基本数据，直接产生了个人隐私可能会泄露出去可能。

（2）冷启动问题。在协同过滤推荐系统刚启动时，或者在一个新用户刚使用系统时，都会遇到因为数据量不够算法产生的推荐不够精准的问题。

（3）稀疏性问题。现实中的数据不像研究数据，它都是很稀疏的，只有很少的用户对很少的项目作给出的评分。

（4）可扩展性问题：协同过滤算法时间复杂度非常高，所以当用户和项目增多的时候，系统压力会成倍的增加，如何再在系统后期仍然保持高效和快速的推荐，也是协同过滤算法中亟待解决的难题。