北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

大数据时代的管理

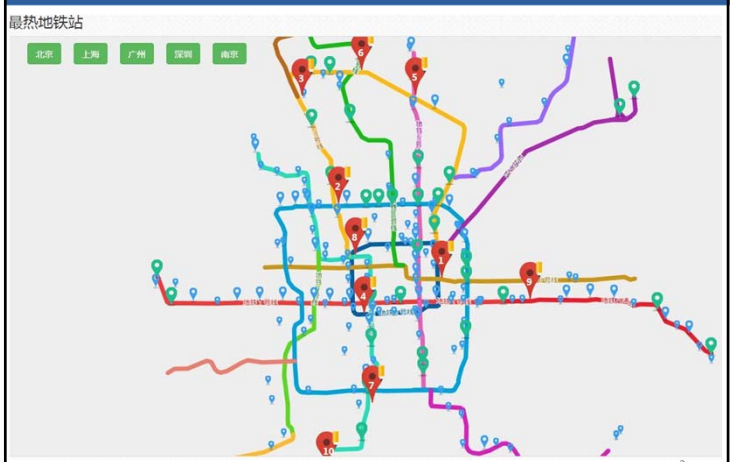
Management in Big Data Era

马宝君 博士 讲师

经济管理学院
电子商务中心
2014年12月22日

1

课前分享1：可视化-最热地铁站



http://www.thinkgis.cn/public/sina/

2

课前分享2：
你点击商户广告后有没有前往实体店，Google知道

- 中国互联网行业发明的“O2O”（Online-to-Offline、线上到线下）这个概念，外国人大多没听说过。但Google在AdWord广告投放系统里新加入的这个功能则正好诠释了这个概念。Google昨天宣布，在AdWord系统里Estimated Total Conversions（预计总转化量）功能当中加入了一个名为“Store Visits”的子项，用于向广告主商户展示广告被点击之后有多少受众前往了实体店。
- 简而言之，商户使用AdWord投放广告的时候，是会登录自己的店铺地理位置信息的，而广告的受众点击广告之后的30天里所产生的地理位置数据，会被AdWord用来和商户的地理位置信息进行比对。从而，Google就能够大概了解到广告的受众最终是否访问了商户的实体店。
- 首先，Google的账号体系使得用户在桌面端和移动端的行为都可以被统一起来；接下来，Google知道用户的浏览数据，看了/点了什么广告，也知道用户使用移动设备的地理位置数据——这些都是用户已经同意让Google获取的信息，Google只是打通了一系列过去相对更分散，没有被整合起来的数据而已。

http://www.pingwest.com/google-adwords-store-visits/

3

上次课程小结：信息检索及信息搜索服务

- **深度业务分析——组合方法及应用**
- **信息检索基础**
  - 基本概念
  - 文档表示、文档权重计算
  - 内容相似度计算
- **链接分析基础**
  - 链接信息的利用
  - PageRank算法的思路、基础模型、扩展模型
  - PageRank的求解
  - 搜索结果的综合排序

4

回顾：深度业务分析——组合方法及应用

- 信息检索及信息搜索服务（文本内容、链接）
- 推荐系统及产品推荐
- 舆情分析及商誉构建（情感）
- 社交网络分析及关系营销
- 用户生成内容（口碑/评论/社交）分析
- ……

5

推荐系统基础

Recommendation Systems Foundations

6

推荐系统基础知识

- 推荐系统出现的背景
- 推荐系统简介
- 推荐系统的模块
  - 用户建模模块
  - 推荐对象建模模块
  - 推荐算法模块
- 推荐系统的评测指标

7

8

推荐系统出现的背景

- 互联网的飞速发展使人们处于一个信息爆炸和信息过载（Information Overload）的时代，新的商业环境在为企业提供商机的同时也提出了新的挑战。
- 如何在虚拟世界中吸引消费者，提高购物满意度并增加客户黏性，成为电子商务平台目前追求的首要目标。
- 面对现阶段海量的信息/数据，对信息的筛选和过滤成为了衡量一个系统好坏的重要指标。

推荐系统与搜索引擎的异同

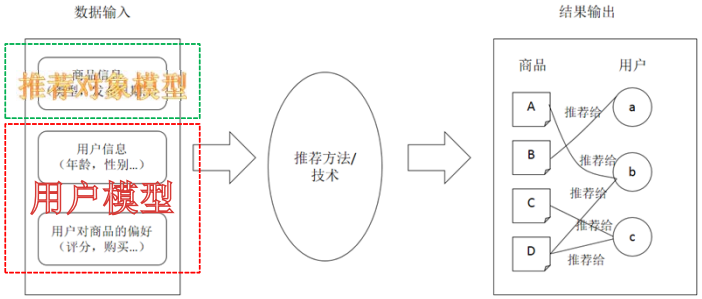
- 相同点
  - 都是帮助用户快速发现有用信息的工具
- 不同点
  - 搜索引擎：需要用户主动提供准确的关键词来寻找信息
  - 推荐系统：不需要用户提供明确的信息需求，而是通过分析用户的历史行为给用户的兴趣建模
- 从某种意义上来说，推荐系统和搜索引擎对于用户来说是两种互补的工具
  - 搜索引擎满足了用户有明确目标时的主动搜索需求
  - 推荐系统能够在用户没有明确目标时帮助他们发现感兴趣的新内容

推荐系统简介

- 推荐系统的任务：联系用户和信息
  - 一方面帮助用户发现对自己有价值的信息
  - 另一方面让信息能够展现在对他感兴趣的人群中，从而实现信息提供商与用户的双赢
- 应用领域广泛
  - 电子商务、电影和视频、音乐、社交网络、阅读、基于位置的服务、个性化邮件和广告等
  - 亚马逊的商品推荐，Facebook的好友推荐，Digg的文章推荐，豆瓣的豆瓣猜，Last.fm和豆瓣FM的音乐推荐，Gmail里的广告等
- 推荐系统充当信息生产者和信息消费者之间的中介
  - 将正确的商品或服务推送给正确的人群

推荐系统的概念

- 推荐系统把用户模型中的兴趣需求信息和推荐对象模型中的特征信息匹配，同时使用想用的推荐算法进行计算筛选，找到用户可能感兴趣的推荐对象，然后推荐给用户



推荐系统是如何工作的

- **以看电影为例**
  - 向朋友咨询：这种方式在推荐系统中成为社会化推荐，即让好友给自己推荐物品
  - 打开搜索引擎，输入自己喜欢的演员名，然后看看返回结果中还有什么电影是自己没有看过的：这种推荐方式在推荐系统中成为基于内容的推荐
  - 查看排行版
  - 找到和自己历史兴趣相似的一群用户，看看他们最近在看什么电影，这种方式被称为基于协同过滤的推荐

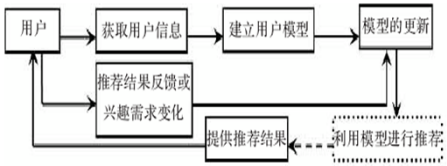
13

推荐系统：用户建模模块

- **一个好的推荐系统要给用户提供个性化的、高效的、准确的推荐，那么推荐系统应能够获取反映用户多方面的、动态变化的兴趣偏好**
- **推荐系统有必要为用户建立一个用户模型，该模型**
  - 能获取、表示、存储和修改用户兴趣偏好，能进行推理
  - 对用户进行分类和识别
  - 帮助系统更好地理解用户特征和类别，理解用户的需求和任务，从而更好地实现用户所需要的功能

14

用户建模的过程



- **获取用户信息：模型输入数据**
  - 用户属性
  - 用户手工输入的信息
  - 用户的浏览行为和浏览内容
  - 推荐对象的属性特征

15

用户模型数据输入的方式

- **显式获取**
  - 用户主动告知：简单直接、全面客观，但很少有用户愿意花时间或不愿意向系统表达自己的喜好
- **隐式获取**
  - 系统通过跟踪用户行为，通过推理学习获取用户的兴趣偏好：减少用户负担，不会打扰用户正常生活，但跟踪的结果未必能准确反映用户的兴趣偏好，且有隐私风险
- **用户的兴趣和需求会随着时间和情景发生变化，用户建模时要考虑到用户长期兴趣偏好和短期兴趣偏好，还要考虑兴趣的变化，目前很多研究关注了用户的长期兴趣，建立了静态模型，用户兴趣更新的动态模型也受到了很多关注，短期兴趣的关注还比较少**

16

用户建模

- **建模的对象有单用户建模和群组建模之分**
  - 单用户建模针对单个用户进行建模，比如基于内容的推荐
  - 群组建模是针对群体用户进行建模，比如协同推荐
- **用户模型的建模方法主要有**
  - 基于机器学习的方法（主要是分类、聚类），例如自动聚类、贝叶斯分类器、决策树归纳和神经网络方法等
  - 遗传算法（Genetic Algorithm）

17

推荐系统：推荐对象建模

- **不同的对象，特征也不相同，目前并没有一个统一的标准来进行统一描述，主要有基于内容的方法和基于分类的方法两大类方法。**
  - 基于内容的方法是从对象本身抽取信息来表示对象，使用最广泛的方法是用加权关键词矢量，该方法通过对一组文档的统计分析得出文档的特征向量。在完成文档特征的选取后，还得计算每个特征的权值，权值大的对推荐结果的影响就大。目前使用最广泛的是TF—IDF方法。
  - 基于分类的方法是把推荐对象放入不同类别中，这样可以把同类文档推荐给对该类文档感兴趣的用户了。文本分类的方法有多种，比如朴素贝叶斯（Naive-Bayes），k最近邻方法（KNN）和支持向量机（SVM）等。对象的类别可以预先定义，也可以利用聚类技术自动产生

18

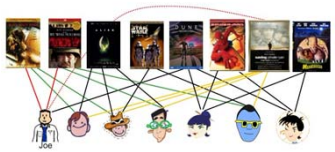
推荐对象建模的一些问题

- 文本等对象特征提取技术相对比较成熟，但推荐系统的对象不一定具有文本特征或文本不足以作为描述，尤其是网络上广泛存在的多媒体数据，自动化的特征提取方法需要结合多媒体内容分析领域的相关技术。
- 推荐系统推荐给用户的对象首先不能与用户看过的对象重复，其次也不能与用户刚刚看过的对象不是太形似或者太不相关，这就是所谓的模型过拟合问题（可扩展性问题）。
- 推荐系统中出现新的对象时，推荐系统尤其是协同过滤系统中，新对象出现后必须等待一段时间才会有用户查看并进行评价，在此之前推荐系统无法对此对象进行分析和推荐，这就是推荐系统研究的另一个难点和重点——冷启动问题。

19

推荐系统：推荐算法模块

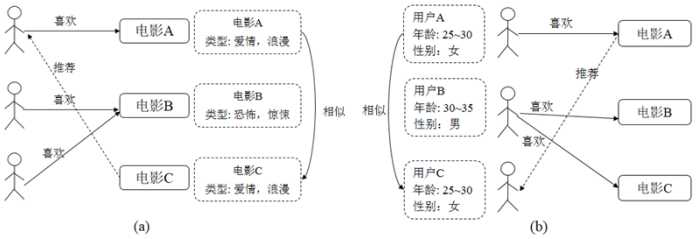
- **推荐算法是整个推荐系统中最核心和关键的部分，在很大程度上决定了推荐系统类型和性能的优劣**
- **公认的推荐算法基本包括以下几种：**
  - 基于内容的推荐
  - 协同过滤推荐
  - 混合推荐



20

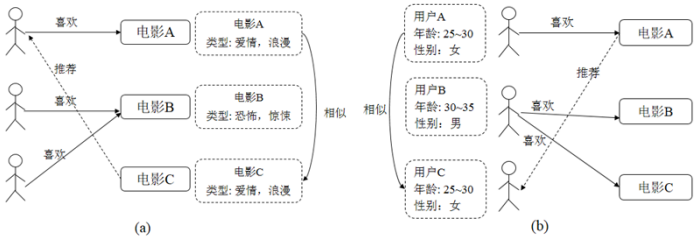
推荐算法1：基于内容的推荐

- 基于内容的推荐方法是指充分利用用户和商品的各种属性特征（如电影的上映时间、演员和题材等，用户的年龄、性别和地理位置等），分析其内容相似度，然后将最相似的商品推荐给用户



推荐算法1：基于内容的推荐

- 基于内容的推荐方法
  - 优点：仅需要使用元数据，而不需要用户与系统的交互即可完成推荐，因此不会遭遇冷启动问题（即：由于缺乏新用户或新产品的历史信息，很难为其进行合适的推荐）
  - 缺点：推荐准确度较低，并且无法挖掘用户的潜在兴趣。



推荐算法2：协同过滤推荐方法

- 协同过滤（Collaborative Recommendation, CF）方法是推荐方法中最经典的方法之一，主要使用用户与系统的交互而形成的反馈数据完成推荐，包括显性反馈（explicit feedback）行为和隐性反馈（implicit feedback）行为
  - 显性反馈行为包括用户明确表示对物品喜好的行为，如评分
  - 隐性反馈行为指的是那些不能明确反应用户偏好的行为，如页面浏览行为
- 主要思想
  - 过去具有相似偏好的用户将来也仍然会具有相似的偏好，是基于历史数据进行推荐的一种方法
  - 对新产品和新用户都会有冷启动的问题，且易受到数据稀疏性的限制

推荐算法2：协同过滤推荐方法

- 协同过滤推荐方法中最主要的是一类称为“基于近邻的推荐方法(Neighborhood-based Recommendation)”
- 基于近邻的推荐方法主要是根据消费者的历史行为数据（如购买历史、关注、收藏行为、打分等）分析用户间或者物品（如：书、电影、音乐、食品、衣物、软件等各种网上商品及服务）间的相似性，从而形成推荐
- 一般可以分为
  - 基于用户的协同过滤方法（UserCF）
  - 基于物品的协同过滤方法（ItemCF）
- 由于电子商务网站中用户的数量一般要远远大于物品的数量，所以UserCF的效率要相对更低一些

推荐算法2.1：基于用户的协同过滤方法

- **UserCF的基本思想**
  - 用户选择某个推荐对象是基于朋友（或近邻）的推荐
  - 也就是说，如果一些用户对某些推荐对象的评分比较相似，则说明这些用户的兴趣偏好相似，那么他们对其他推荐对象的评分应该也是相似的
  - 首先找到和目标用户兴趣偏好相似的最近邻居，然后根据他的最近邻居对推荐对象的评分来预测目标用户对未评分的推荐对象的评分，选择预测评分最高的若干个推荐对象作为推荐结果反馈给用户

$$Sim(u,v) = \frac{\sum_{i\in\mathcal{I}_{uv}}(r_{ui}-\bar{r}_u)(r_{vi}-\bar{r}_v)}{\sqrt{\sum_{i\in\mathcal{I}_{uv}}(r_{ui}-\bar{r}_u)^2\sum_{i\in\mathcal{I}_{uv}}(r_{vi}-\bar{r}_v)^2}} \qquad \hat{r}_{ui} = \bar{r}_u + \frac{\sum_{v\in NNs_u} sim(u,v)\times(r_{vi}-\bar{r}_v)}{\sum_{v\in NNs_u}(|sim(u,v)|)}$$

25

推荐算法2.2：基于物品的协同过滤方法

- **ItemCF的基本思想**
  - 基于用户对推荐对象品牌的信任而进行的推荐
  - 如果大部分用户对一些推荐对象的评分比较相似，则当前用户对这些项的评分也比较相似。就好像很多用户对某个品牌比较信任，则其他用户就比较容易选择该品牌的产品
  - 首先找到目标对象的最近邻居，由于当前用户对最近邻居的评分与对目标推荐对象的评分比较类似，所以可以根据当前用户对最近邻居的评分预测当前用户对目标推荐对象的评分，然后选择预测评分最高的若干个目标对象作为推荐结果呈现给当前用户

$$Sim(i,j) = \frac{\sum_{u\in\mathcal{U}_{ij}}(r_{ui}-\bar{r}_i)(r_{uj}-\bar{r}_j)}{\sqrt{\sum_{u\in\mathcal{U}_{ij}}(r_{ui}-\bar{r}_i)^2\sum_{u\in\mathcal{U}_{ij}}(r_{uj}-\bar{r}_j)^2}} \qquad \hat{r}_{ui} = \bar{r}_i + \frac{\sum_{j\in NNs_i} sim(i,j)\times(r_{uj}-\bar{r}_j)}{\sum_{j\in NNs_i}(|sim(i,j)|)}$$

26

推荐算法3：混合推荐方法

- **由于不同的推荐方法都有各自的优缺点，在实际应用中大部分推荐系统通常是混合使用各种推荐方法，达到扬长避短的目的，从而产生更符合用户需求的推荐**
- **具体的混合方法主要包括**
  - 加权型混合推荐方法（将来自不同推荐方法生成的候选结果进一步组合加权）
  - 转换型混合方法（采取一定的标准在不同的推荐方法之间变换，以达到更高的预测准确率）
  - 瀑布型混合推荐方法（将不同的推荐方法视为不同粒度的过滤器，前一个方法的输出作为后一个方法的输入）
  - ……

27

推荐系统的评测指标

- **直观评估原则**
  - 好的推荐系统应该在推荐准确的基础上，给所有用户推荐的物品尽量广泛，给单个用户推荐的物品尽量覆盖多个类别，同时不要给用户推荐太多热门物品
- **具体来说，推荐系统存在3个参与方：用户、物品提供者和网站平台。因此在评测推荐方法时，应该同时考虑三方的利益，实现所有参与者的共赢**
  - 首先需要满足用户的需求，给用户推荐那些令他们感兴趣的商品；
  - 其次要尽量使各个供应商的产品都能被推送给合适的用户，而不是仅仅推荐畅销产品；
  - 最后，好的推荐系统应能够让系统本身收集到高质量的用户反馈，不断完善推荐的质量

28

推荐系统的评测指标

- 用户满意度
- 预测准确率
- 覆盖率
- 多样性
- 新颖性
- 惊喜度
- 信任度
- 实时性
- 鲁棒性
- ……

29

用户满意度

- 用户作为推荐系统的参与者，其满意度是评测推荐系统的最重要指标
- 用户调查获得用户满意度主要是通过调查问卷的形式
- GroupLens曾经做过一个论文推荐系统的调查问卷，请问下面哪句话最能描述你看到推荐结果的感受？
  - 推荐的论文都是我非常想看的
  - 推荐的论文很多我都看过了，确实是符合我兴趣的不错论文
  - 推荐的论文和我的研究兴趣是相关的，但我并不喜欢
  - 不知道为什么会推荐这些论文，它们和我的兴趣丝毫没有关系

30

用户满意度（续）

- 在线系统中，用户满意度主要通过一些对用户行为的统计得到
  - 电子商务网站中，用户如果购买了推荐的商品，就表示他们在一定程度上满意，可以利用购买率度量用户的满意度
  - 有些网站会通过设计一些用户反馈界面收集用户满意度，有对推荐结果满意或者不满意的反馈按钮，统计两种按钮的单击情况
  - 更一般的情况下，我们可以用点击率、用户停留时间和转化率等指标度量用户的满意度

31

预测准确率：Accuracy

- 预测准确率是衡量一个推荐系统质量最重要的指标，是用来度量一个推荐方法预测用户行为的能力，一般通过离线实验来计算。常用的标准数据集包括：Netflix大赛数据集、MovieLens数据等。
- 首先将离线数据集分成训练集和测试集，然后通过在训练集上建立用户的偏好模型来预测在测试集上的行为，最后计算预测行为和实际行为的重合度作为预测准确率。
- 对推荐方法的评测通常有两个方向：评分预测和排序预测，二者分别使用不同的测度。

32

预测准确率：Accuracy

- **评分预测**
  - 均方根误差（Root Mean Square Error，RMSE） $RMSE=\sqrt{\frac{\sum_{u,i\in T}(r_{ui}-\hat{r}_{ui})^2}{|T|}}$
  - 平均绝对误差（Mean Absolute Error, MAE） $MAE=\frac{\sum_{u,i\in T}|r_{ui}-\hat{r}_{ui}|}{|T|}$
- **排序预测（Top-k推荐）：信息检索领域的传统测度**
  - 查全率 $recall=\frac{\sum_{u\in U}|r(u)\cap T(u)|}{\sum_{u\in U}|T(u)|}$
  - 查准率 $precision=\frac{\sum_{u\in U}|r(u)\cap T(u)|}{\sum_{u\in U}|r(u)|}$

33

覆盖率和多样性

- **覆盖率（Coverage）**
  - 是内容提供商较为关注的指标，它描述了一个推荐系统对网站平台上长尾商品的挖掘能力（此外还有：信息熵和基尼系数）

$$coverage=\frac{|\sum_{u\in U}R(u)|}{|I|}$$

- **多样性（Diversity）**
  - 描述了推荐结果中商品两两之间的不相似性，以覆盖用户多个方面的兴趣点，增加用户找到感兴趣产品的概率
  - 多样性推荐列表的好处用一句俗话表述就是“不在一棵树上吊死”

$$diversity(R(u))=1-\frac{\sum_{i,j\in R(u),i\neq j}sim(i,j)}{\frac{1}{2}|R(u)||R(u)-1|}\qquad diversity=\frac{1}{|U|}\sum_{u\in U}diversity(R(u))$$

34

新颖性

- **新颖的推荐是指给用户推荐那些他们以前没有听说过的物品**
- **在一个网站中实现新颖性的最简单方法：把那些用户之前在网站中对其有过行为的物品从推荐列表中过滤掉**
- **O'scar Celma在博士论文中研究了新颖度的评测**
  - 评测新颖度的最简单方法是利用推荐结果的平均流行度，因为越不热门的物品越可能让用户觉得新颖
  - 如果推荐结果中物品的平均热门程度较低，那么推荐结果就可能有比较高的新颖性
  - 但是，用推荐结果的平均流行度度量新颖性比较粗略，因为不同用户不知道的东西是不同的
- **要准确地统计新颖性需要做用户调查**

35

惊喜度

- **惊喜度是最近这几年推荐系统领域最热门的话题**
- **惊喜度和新颖性是有区别的：**
  - 如果推荐结果和用户的历史兴趣不相似，但却让用户觉得满意，那么就可以说推荐结果的惊喜度很高
  - 推荐的新颖性仅仅取决于用户是否听说过这个推荐结果
- **提高推荐惊喜度需要提高推荐结果的用户满意度，同时降低推荐结果和用户历史兴趣度的相似度**

36

信任度

- 度量推荐系统的信任度只能通过问卷调查的方式，询问用户是否信任推荐系统的推荐结果
- 提高推荐系统的信任度的两种方法：
  - 增加推荐系统的透明度（如，提供推荐解释）
  - 考虑用户的社交网络信息，利用用户的好友信息给用户做推荐，并且用好友进行推荐解释

37

实时性

- 物品（新闻、微博等）具有很强的时效性，需要在物品还具有时效性时就将它们推荐给用户
- 推荐系统的实时性包括两个方面：
  - 推荐系统需要实时地更新推荐列表来满足用户新的行为变化
  - 推荐系统需要能够将新加入系统的物品推荐给用户

38

鲁棒性（健壮性）

- 任何一个能带来利益的算法系统都会被人攻击，这方面最典型的例子就是搜索引擎（作弊和反作弊斗争）
- 推荐系统目前也遇到了同样的作弊问题，而鲁棒性指标衡量了一个推荐系统抗击作弊的能力
- 最著名的作弊方法：行为注入攻击（profile injection attack）
- 算法鲁棒性的评测主要利用模拟攻击
- 在实际系统中，提高系统的鲁棒性，除了选择鲁棒性高的算法，还有以下方法：
  - 设计推荐系统时尽量使用代价比较高的用户行为
  - 在使用数据前，进行攻击检测，从而对数据进行清理

39

小结：推荐系统基础知识

- 推荐系统出现的背景
- 推荐系统简介
- 推荐系统的模块
  - 用户建模模块
  - 推荐对象建模模块
  - 推荐算法模块
- 推荐系统的评测指标

40

期末课程论文说明

- **主题要求**
  - 必须与“大数据管理”相关
  - 建议围绕所学专业背景下的“大数据管理问题”展开
- **内容要求**
  - 不少于4000字，版式：word中正文小四字体，1.5倍行距
  - 独立完成，不得大段拷贝或直接引用网上、书上及他人已发布内容，需要适当引用时请在引用位置注明参考文献来源（查重）
  - 论文内容框架（建议）：
    - 1. 学习本课程的心得体会、感受，对本课程教学的建议和意见（必有）
    - 2. 论文背景介绍
    - 3. 论文涉及的大数据问题及管理需求、策略和意义（可举实例说明）
    - 4. 本人对该大数据问题的看法、观点及讨论
    - 5. 总结
    - 6. 参考文献和资料

41

期末课程论文说明（续）

- **论文提交要求**
  - 需要以电子版提交，建议提交word版本
  - 作业提交邮箱：bigdata_homework@163.com
  - 作业提交截止时间：**第19周周日（2015.01.11）24时**
- **其他说明**
  - **邮件标题和电子版论文文件请务必按照“学号_班级_姓名.docx”命名，例如“2014211234_2014212103_张三.docx”，也请在邮件中留下姓名、学号及联系方式，以备论文有问题时能够联系到；**
  - 请在截止时间之前提交论文（不要在截止时间附近，以避免系统原因过期），过期将不再接收论文提交，成绩为0，请务必注意；
  - 每次提交论文后，作业邮箱都会有“已收到邮件”的自动回复，如未收到自动回复，表示发送不成功，请在截止时间内重新提交；
  - 论文评分的关注重点
    - 有效的课程建议和意见
    - 关注问题的新颖度
    - 个人分析和讨论的深度
    - 论文的整体工作量

42