新媒体大数据分析

New Media Big Data Analysis

第二章 数据分析

黄振亚，朱孟潇，张凯

课程主页：
http://staff.ustc.edu.cn/~huangzhy/Course/NM2024.html

助教：陈宗阳
bigdata_2024@163.com

数据预处理

- 大数据环境下的数据特征
- 为什么需要进行预处理
- **预处理的基本方法**
  - 数据清理
  - **数据集成**
  - 数据变换
  - 数据规约

数据预处理：数据集成

数据的相关性分析

- **无序数据：每个数据样本的不同维度是没有顺序关系的**
  - 余弦相似度、相关度、欧几里得距离、Jaccard
- **有序数据：对应的不同维度(如特征)是有顺序(rank)要求的**
  - 在信息检索中，如何判断不同检索方法返回的页面序列的优劣
  - 在推荐系统中，如何判断不同推荐序列的好坏
    - **Spearman Rank(斯皮尔曼等级)相关系数**
    - **归一化的折损累计增益(NDCG)**
    - 肯德尔相关性系数
      - kendall correlation coefficient
- 课外阅读：PageRank算法

i	相关度
1	3
2	3
3	2
4	0
5	1
6	2

方法返回结果

i	相关度
1	3
2	3
3	2
4	2
5	1
6	0

真实结果

数据预处理：数据集成

数据的相关性分析—举例

- 已知：6个网页的相关度是3，2，3，0，1，2，所以在信息检索中，最好的返回结果应当如(a)所示。
- 如果我们设计了两个检索算法，返回结果分别是(b)和(c)，请问哪个方法的结果与真实结果更相似？

i	相关度
1	3
2	3
3	2
4	2
5	1
6	0

(a)真实结果

i	相关度
1	3
2	3
3	0
4	2
5	2
6	1

(b)方法1返回结果

i	相关度
1	3
2	3
3	2
4	0
5	2
6	1

(c)方法2返回结果

数据预处理：数据集成

- 有序数据的距离度量(信息检索、推荐系统等)
  - Spearman Rank(斯皮尔曼等级)相关系数
    - 比较两组变量的相关程度
    - 当关系是非线性时，它是两个变量之间关系评价的更好指标

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

    - ρ_s：表示斯皮尔曼相关系数
    - d_i^2：表示每一对样本之间等级的差
    - n：表示样本容量
  - ρ_s的范围：-1 to 1 (正相关:$\rho_s > 0$，负相关:$\rho_s < 0$，不相关:$\rho_s = 0$)

数据预处理：数据集成

- 有序数据的距离度量(信息检索、推荐系统等)
  - **Spearman Rank(斯皮尔曼等级)相关系数**

  - X = (a, b, c, d, e, f)
  - Y = (c, a, e, d, f, b)

  $$d_i = Y_i - X_i$$

  - $d_i^2 = (4, 1, 4, 0, 1, 16)$

  $$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

  - $\rho = 1 - \frac{6(26)}{6(36-1)} \approx 1 - 0.743 = 0.257$

数据预处理：数据集成

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

数据的相关性分析——练习题2 （计算Spearman）

- 已知6个网页的相关度是3，2，3，0，1，2，所以在信息检索中，最好的返回结果应当如(a)所示。如果我们设计了两个检索算法，返回结果分别是(b)和(c)，
- 请问：哪个方法的结果与真实结果更相似？

i	相关度
1	3
2	3
3	2
4	2
5	1
6	0

(a)真实结果

i	相关度
1	3
2	3
3	0
4	2
5	2
6	1

(b)方法1返回结果

i	相关度
1	3
2	3
3	2
4	0
5	2
6	1

(c)方法2返回结果

只考虑了每个位置的数据与真实数据的顺序差异，但是没有考虑到不同位置(position)的重要性差异

数据预处理：数据集成

- 有序数据的距离度量(信息检索、推荐系统等)
  - NDCG(Normalized Discounted cumulative gain)
    - CG(累计增益)：只考虑到了相关性的关联程度，没有考虑每个推荐结果处于不同位置对整个推荐效果的影响

$$CG_k = \sum_{i=1}^{k} rel_i$$

rel_i表示处于位置 ii 的推荐结果的相关性

- DCG(折损累计增益)：就是在每一个CG的结果上处以一个折损值，目的就是为了让排名越靠前的结果越能影响最后的结果

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i}-1}{log_2(i+1)}$$

- i表示推荐结果的位置， i越大，则推荐结果在推荐列表中排名越靠后推荐效果越差，DCG越小

Qi Liu, Yong Ge, Enhong Chen, and Hui Xiong. Personalized Travel Package Recommendation. ICDM'2011, **(Best Research Paper Award)**

数据预处理：数据集成

- **有序数据的距离度量(信息检索、推荐系统等)**
  - **NDCG(** Normalized Discounted cumulative gain **)**
    - **NDCG：**由于搜索结果随着检索词的不同，返回的数量不一致，而DCG是一个累加的值，没法针对两个不同的搜索结果进行比较，因此需要标准化处理，这里是除以IDCG:

$$NDCG_k = \frac{DCG_k}{IDCG_k}$$

IDCG为理想（ideal）情况下最大的DCG值，指推荐系统为某一用户返回的最好推荐结果列表(或者，真实的数据序列)

数据预处理：数据集成

$$NDCG_k = \frac{DCG_k}{IDCG_k}$$

- 例，假设一个推荐系统为用户推荐了3部电影，顺序为A，B，C，用户实际对这三部电影的偏好为B > A > C, 假定A，B，C三部电影的相关性分数分别为2，3，1，那么对于系统返回的结果有：
  - CG@3 = 2 + 3 + 1= 6
  - DCG@3 = 3 + 4.42 + 0.5 = 7.92

$$CG_k = \sum_{i=1}^{k} rel_i$$

- 理想情况下，系统给出的电影排序应该为B，A，C
  - IDCG@3 = 7 + 1.89 + 0.5 = 9.39
- 可以计算NDCG@3
  - NDCG@3 = 7.92 / 9.39 = 0.84

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i}-1}{log_2(i+1)}$$

i	movie	rel	$\frac{2^{rel_i}-1}{log_2(i+1)}$
1	A	2	3
2	B	3	4.42
3	C	1	0.5

方法返回结果

i	movie	rel	$\frac{2^{rel_i}-1}{log_2(i+1)}$
1	B	3	7
2	A	2	1.89
3	C	1	0.5

真实结果

课堂练习：数据集成

$$CG_k = \sum_{i=1}^{k} rel_i$$

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i}-1}{log_2(i+1)}$$

$$NDCG_k = \frac{DCG_k}{IDCG_k}$$

数据相关性分析——练习题3

- 已知6个网页的相关度是3，2，3，0，1，2，所以在信息检索中，最好的返回结果应当如(a)所示。如果我们设计了两个检索算法，它们的返回结果分别是(b)和(c)，请问哪个方法的结果与真实结果更相似（根据**NDCG的计算结果**）。

i	相关度
1	3
2	3
3	2
4	2
5	1
6	0

(a)真实结果

i	相关度
1	3
2	3
3	0
4	2
5	2
6	1

(b)方法1返回结果

i	相关度
1	3
2	3
3	2
4	0
5	2
6	1

(c)方法2返回结果

可以只列出计算公式，不用给出计算结果

- **0.9746**
- **0.9889**

数据预处理：数据集成

课后阅读

- Defu Lian, Haoyu Wang, Enhong Chen, Xing Xie. LightRec: a Memory and Search-Efficient Recommender System. WWW 2020.
- Qi Liu, Zhenya Huang, Enhong Chen,, EKT: Exercise-aware Knowledge Tracing for Student Performance Prediction, TKDE
- Zhenya Huang, Qi Liu, Enhong Chen, et al, Question Difficulty Prediction for READING Problems in Standard Tests, AAAI'2017
- Qi Liu, Yong Ge, Enhong Chen, and Hui Xiong. Personalized Travel Package Recommendation. ICDM'2011, (Best Research Paper Award)
- PageRank算法