# 公益基金会推荐系统

## 团队名称：C3

### 1. 摘要：

本方案结合项目型基金会固有的特征和资助型基金会对项目型基金会偏好（相互关系），使用协同过滤的方法，为资助型基金会推荐适合他们投资的项目型基金会。同时，新的基金会在输入相应信息之后，根据基本特征相近的办法，也可得到合适的推荐。

### 2. 背景：

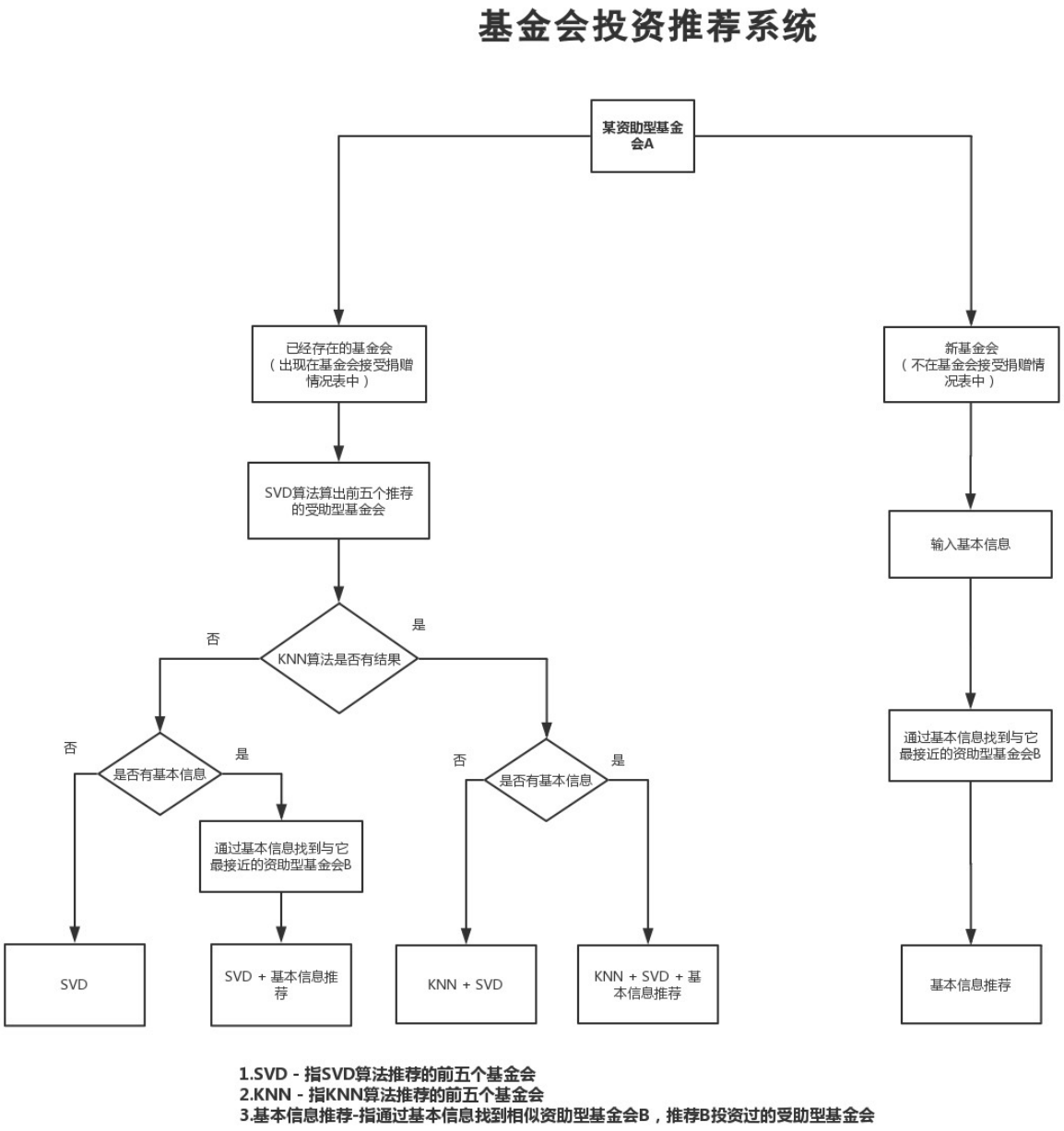
不同的公益基金会由于自身建立的目标，基金会财务状况，甚至一些地理位置带来的“地缘”因素，会对投资的项目型基金会产生不同的偏好。资助型基金会通过他们对项目型基金会的合理投资，不仅更高效地将资本和执行方结合实现了自身的愿景，也对社会资源进行有效合理的分配。他们的作用，是在政府无法管控到或管控不足的领域内，由民间力量聚集起来的看不见的手，为社会问题提供了有效的解决方案。在基金会中心网在案的 6125 家基金会记录中，有 3813 家基金会都曾经投资过其他的基金会，占总数的 62.29%。因而，解决这一问题具有深远的社会影响。

### 3. 需求：

我们的系统可以满足两类资助型基金会的需求。

**对于已经纳入基金会中心网“基金会接受捐赠情况”的资助型基金会**，我们可以通过它**历史投资**的数据，为它推荐符合它便好的项目型基金会。

**对于未纳入基金会中心网“基金会接受捐赠情况”的资助型基金会，我们无法得知它历史投资的数据**。但是我们会通过问卷的方式，要求基金会简要回答 10 余项关于基本信息的问题。通过它的基本特征，为它寻找到和它最相似的已经在基金会中心网数据库中的基金会，进而做出推荐。




**4. 最终方案及数据来源：**

**最终方案：**

**a. 对于已经纳入基金会中心网“基金会接受捐赠情况”的资助型基金会推荐系统的设计，我们主要采纳了 SVD 和 KNN 算法。**

SVD 算法，即是 Netflix 推荐系统中涉及的经典推荐方法。由于数据本身对应关系比较少，大部分投资型基金会只投资了一个受助型基金会，用 SVD 算法得到的结果集中在被投资多次

的“热门”项目型基金会，我们结合了 KNN 算法 。KNN 基于资助型基金会对项目型基金会的打分进行了系统过滤，当中相似度的概念我们选取了 Cosine 距离 。两个算法我们都是都过 R 的”Surprise”包实现，分别取前 5 的结果（如果可以有结果）。

**关于 SVD 和 KNN 算法用到的打分的数据清洗和具体计算如下：**

**数据清洗：**

首先，通过对基金中心网“基金会接受捐赠情况”记录的爬虫，我们得到 2011 年到 2015 年的 32550 条数据；

其次，筛选出当中捐款方为基金会的，得到 32550 条数据；

再者，我们计算出 2011 年到 2015 年投资次数大于等于 1 的基金会共计 2010 家；对于有多次投资的基金会，我们选取 2011-2015 年间投资金额的平均值，筛选出有财务数据（净资产）的企业共计 4306 家；

最后，通过结合财务信息数据和捐赠情况数据，并对年份去重，我们最终得到 1564 个基金会数据。

*数据清理使用软件：stata14 （详见 DataClean 文件夹）*

*请将输入数据和代码置于同一路径文件夹下，并在代码开头将 cd 后的内容修改为该路径*

*输入数据：*

*FundationList_final.xls 网络爬虫的 32550 条捐赠数据*

*Foundation_asset_2013.xlsx 2013 年的财务数据*

*Foundation_asset_2014.xlsx 2014 年的财务数据*

*Foundation_asset_2015.xlsx 2015 年的财务数据*

*输出数据：*

*"score_all.xls " 最终评分数据：其中 user 和 item 分别为资助型基金会和项目型基金会，各自对应相应的 userid 和 itemid。rate_t 为投资次数，rate_a 为投资金额占净资产的百分比，score_t 为投资次数打分，score_a 为投资金额打分，score_all 为最终得分。*

**打分参考了两个指标**：

**投资额占净资产的百分比**：由于基金会本身财务状况不同，对项目型基金会的投资额自然会产生影响。我们通过基金会的净资产去除了这一影响，更加公允地表示了基金会投资意愿。我们的给分标准是：前 10%是 10 分，10%-20%是 20 分，以此类推到 100 分。对极端值的处理：<0.01 的是 5 分， >100 的是 100 分。

**投资次数（2011-2015 年之间）**：投资次数带有强烈的偏好暗示，因此我们在下面的最终给分中给了投资次数更大的权重。对于投资次数这一部分而言，我们的给分标准是：31 分为基地，多增加一次投资+3 分，最高 100 分（投资了 24 次）。

最重的分数 Score_all = 投资额占净资产的百分比的得分* 30% +投资次数的得分* 70%，分布如下：

score_all

| Percentiles | | Smallest | | |
|---|---|---|---|---|
| 1% | **24.7** | **23.2** | | |
| 5% | **24.7** | **23.2** | | |
| 10% | **27.7** | **23.2** | Obs | **1,564** |
| 25% | **32.8** | **23.2** | Sum of Wgt. | **1,564** |
| 50% | **39.7** | | Mean | **40.38344** |
| | | Largest | Std. Dev. | **9.611228** |
| 75% | **48.7** | **78.1** | | |
| 90% | **51.7** | **79** | Variance | **92.3757** |
| 95% | **53.8** | **80.5** | Skewness | **.2328002** |
| 99% | **62.5** | **82.9** | Kurtosis | **3.0026** |

**b. 对于未纳入基金会中心网“基金会接受捐赠情况”的资助型基金会推荐系统的设计，我们通过关键特征值，找到和它最相似的纳入系统的基金会，推荐最相似基金会所投过的项目型基金会。**

当前面一种情况做完之后，我们面临一个新的问题，新的资助型基金会（未纳入基金会中心网“基金会接受捐赠情况“的资助型基金会）出来了怎么办？由于我们没有任何关于他的投资信

息，我们可以通过用基金会本身的基本信息找到已经存在在基金会中心网官方资料中离它最近的基金会，把这个已经存在的资助型基金会的投资过的项目型基金会推荐给这个新的基金会。我们会当探索到用户输入的基金会为新的资助型基金会时，询问 17 个关键指标，利用这些特征值根据已有的数据进行相似度计算。这 17 个关键指标是我们整合了资助型基金会所有的基本信息，财务信息和透明度打分，运用 PCA 的方法提取了 17 个重要的特征值 。

补充一点，相信聪明的您已经从我们的流程表中看出，当已经存在的基金会（前面一种情况所讨论的）同样出现在官方那个指标的表格里，我们会帮它找到 最近的资助型基金会，然后把这个基金会投过的项目型基金会推荐给它。

**整体数据来源补充说明：**

由于数据量的大小对推荐系统的精准度有显著的影响，我们通过爬虫取得了基金会中心网上 6125 家基金会的全部记录 32 万多条投资信息，而没有直接采用比赛组委会给出的 605 家基金会的数据。

爬虫的代码在 spider 文件下：invest_inf.py 爬取了基金会中心网 “基金会接受捐赠情况”所有资助型基金会对项目型基金会的投资情况数据；invest_url.py 爬取了所有在 invest_info.py 列表里的投资型基金会的详细地址；finance_info.py 是根据每个投资型基金会详细地址，爬取了他们 2013 年度到 2015 年度的财务数据。

## 5. 研究方法

本方案结合项目型基金会固有的特征和资助型基金会对项目型基金会偏好（相互关系），通过这两个维度的信息， 为资助型基金会推荐适合他们投资的项目型基金会。

## 6. 研究用例

1. 代码需要在 python3 的环境下运行。需要安装一些 python 的包，具体参考 requirements.txt。也可以一键安装所需要的环境： pip install -r requirements.txt。
2. 在 shell 中执行 main.py。
3. 我们以资助型公益基金会的用户视觉视角，可以看到具体如下基本流程：

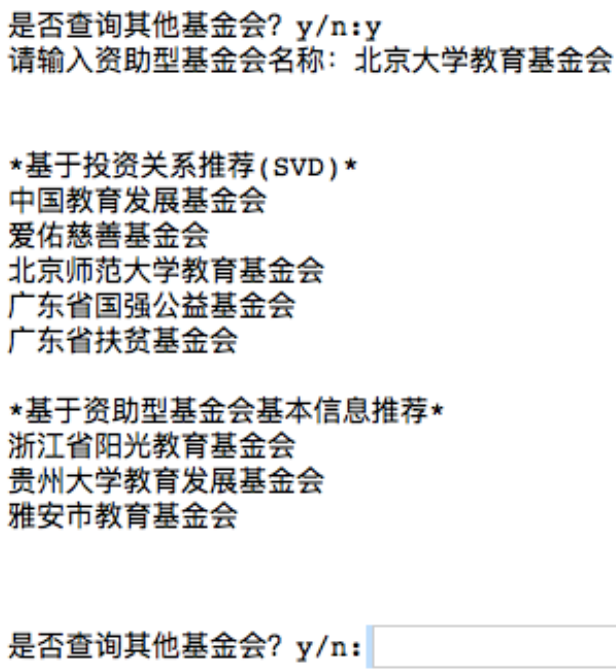
a.首先，系统询问是否已经装好“surprise”这个 Python 的包。如果没有装好，可以输入“2”，查询时系统会从已经存在本次的查询结果里面提取。装好”surprise”这个包的用户，输入“1”，系统计算完成 Cosine 相似度后，就会跳出“请输入资助型基金会名称”（大概需要 2-3 秒时间）。

```
请选择：
1.我已经安装好surprise包，重新跑一次模型
2.利用已经存在本地数据库
(请输入数字):1
Computing the cosine similarity matrix...
Done computing similarity matrix.

请输入资助型基金会名称：
```

b. 我们在“请输入资助型基金会名称”处输入“爱佑慈善基金会，得到如下查询结果，是”已经存在的基金会中，既有 KNN 结果又有基本信息“的情况，得到了“KNN+SVD+基金信息”的推荐。

```
请输入资助型基金会名称：爱佑慈善基金会


*基于投资关系推荐(KNN)*
北京凯恩克劳斯经济研究基金会
海南省医疗救助基金会
陕西师范大学教育基金会
上海市华侨事业发展基金会
北京感恩公益基金会

*基于投资关系推荐(SVD)*
中国教育发展基金会
爱佑慈善基金会
广东省国强公益基金会
北京师范大学教育基金会
清华大学教育基金会

*基于投资型基金会基本信息推荐*
传媒大学教育基金会
辽宁省红十字基金会
重庆市儿童医疗救助基金会
陕西省红十字基金会


是否查询其他基金会? y/n:
```

c.系统跳出，“是否查询其他基金会”，我们输入”y”,继续查询“北京大学教育基金会”。这是一个“已经存在的基金会中，没有 KNN 结果但是有基本信息”的样例。输出的结果是“SVD+基本信息”的推荐。




d. 我们在“是否查询其他基金会”中输入”y”,继续查询“北京交通大学教育基金会”。我们发现系统跳出“该基金会不在数据库中/或者不是资助型基金会，请输入相关信息方便我们为你匹配对应的受助型基金会！”,并随之要求输入 17 项基本信息。（如图输入信息为虚构）。输入全部信息后，系统根据算法给出了基于资助型基金会本身基本信息的推荐。

```
是否查询其他基金会? y/n:y
请输入资助型基金会名称：北京交通大学教育基金会


该基金会不在数据库中/或者不是资助型基金会，请输入相关信息方便我们为你匹配对应的受助型基金会!
捐赠收入(元):100000
总收入(元):200000
境内捐赠(元):50000
其他收入(元):50000
自然人捐赠(元):0
公益事业支出(元):10000
业务活动成本(元):2000
总支出(元):17000
全职员工:36
评估等级是否为"未参评"(是请输1，否请输0):0
行政办公支出(元):4000
基金会评估等级是否为"5A"(是请输1，否请输0):0
成立时间(YYYY-MM-DD):1990-01-01
管理费用(元):4500
透明度分数:90
行业领域是否包括"心理健康"(是请输1，否请输0):0
行业领域是否包括"扶贫助困"(是请输1，否请输0):0


*基于资助型基金会基本信息推荐*
北京中间艺术基金会


是否查询其他基金会? y/n:
```

**7. 讨论**

a. 基金会这个案例模式和传统的商品购买/电影评分推荐系统的分析存在一定的数据差距。传统的商品购买/电影评分推荐系统，一个客户会购买多个商品（一个观众会看多部电影）；同时，一个商品客户购买（一部电影会被多个观众观看），是具有丰富数据量的典型多对多。但是对于基金会现有的数据库来看，很多资助型基金会只投资过一个项目型基金会，很多项目型基金会也只被一个资助型基金会投资过。这会很大程度上影响推荐系统的可靠程度。

b. 本方案暂时使用了项目型基金会固有的特征和资助型基金会对项目型基金会偏好（相互关系）这两个维度的信息。未来可以考虑将项目型基金会的固有特征加到模型中，使得推荐更加完善和个性化。

c. 由于我们使用的算法需要大量的数据作为支撑，我们在投资信息上，选择爬虫的方式获取了共计 32550 条数据，相比于官方给出的 3200 条数据，为 a 类情况的资助型基金会研究极大地增加了样本量。但在 b 类情况的资助型基金会研究中，我们仅使用了官方给出的 605 数据作为支撑，未来可以考虑进行更全面的数据爬虫，为符合 b 类情况的资助型基金会，提供更完整的分析。