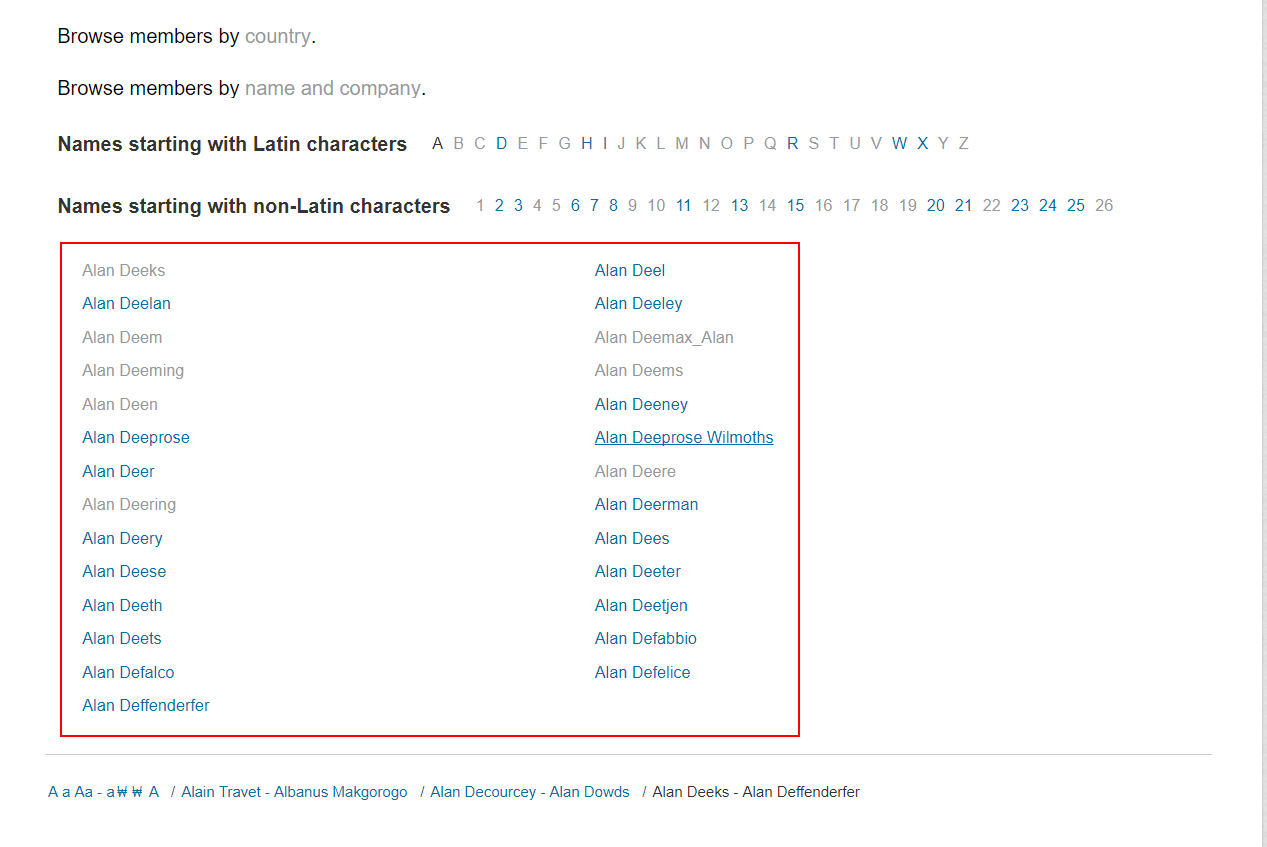
1. 解决下图中的翻页问题，从类似于<http://www.linkedin.com/directory/people-a-23-23-2>的页面进去具体的项目时，如下图1

图1.

，一些会直接进入到profile页面，有一些会进入到如图2的一个列表，大部分只有一页，极个别列表有多页，对图2的页面进行处理，得到其中的user url，然后根据user url再进行profile的爬取。

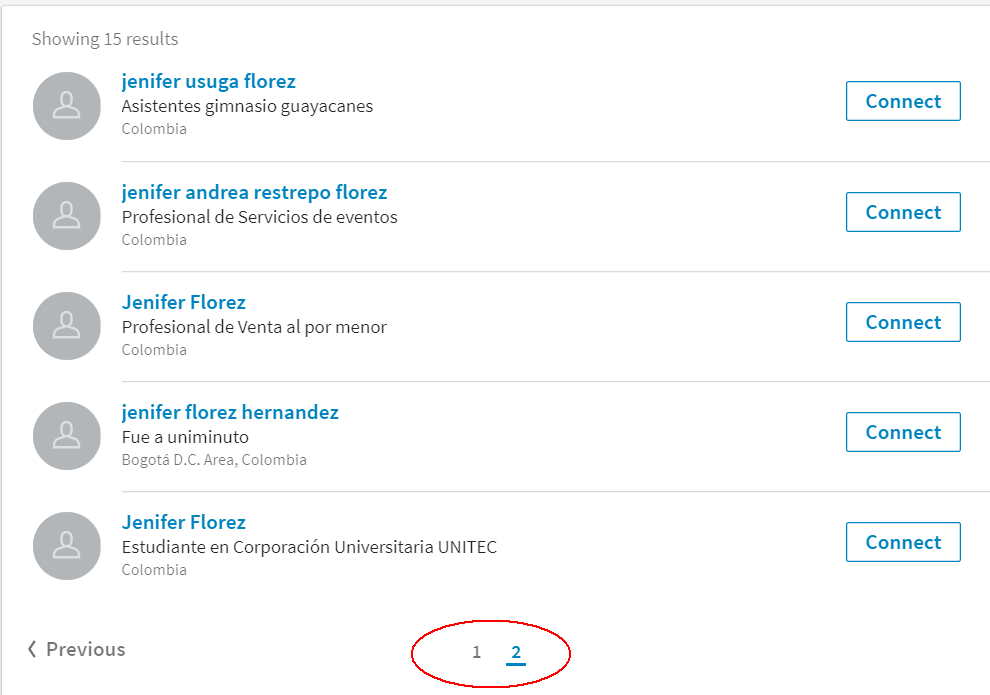


图2.

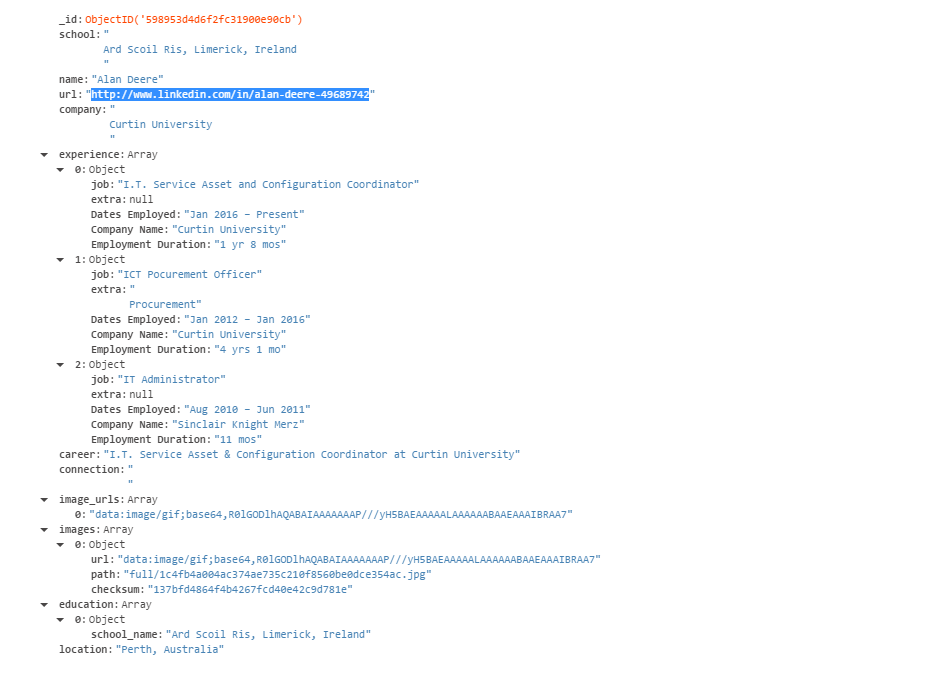
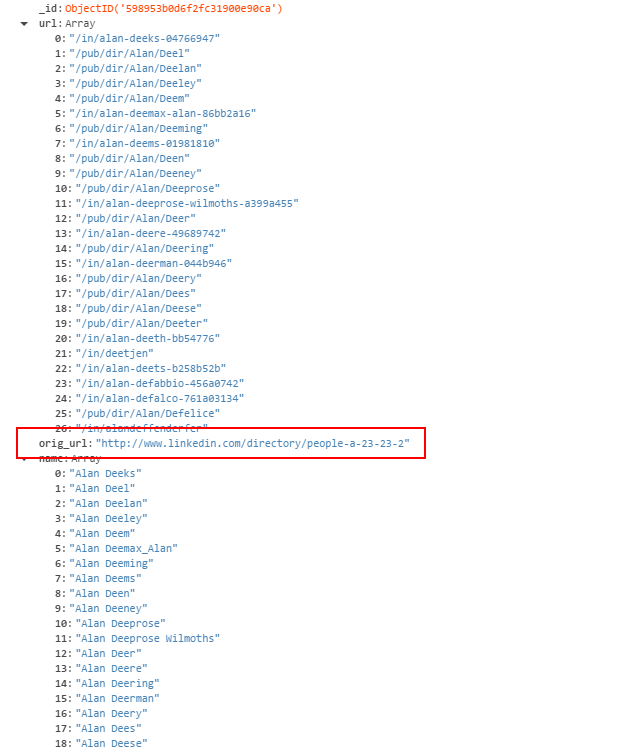
2. 数据库中存贮的数据主要包含3个表，profiles，scrapy\_items，scrapy\_items\_failed。profiles表中主要存放profile数据如图3. 

图3.

scrapy\_items表主要存放已经爬取过的类似于<http://www.linkedin.com/directory/people-a-23-23-2>的数据，如图4. 其中的url数组为相应url（红色方框）页面中包含的用户名url地址，name数组为相应姓名对应的数组。

Scrapy\_items\_failed主要存放因为各种原因导致爬取失败的url的数据，格式类似图4.

3. 爬虫在启动时会根据数据库中的数据判断url是否已经爬取过，跳过已经爬取的url，这样来保证爬虫多次运行的情况下，不会爬取重复数据，也能够保证数据库数据的唯一性。