链滴

中国娱乐圈关系挖掘

作者：xjtushilei
原文链接：https://ld246.com/article/1487580002863
来源网站：链滴
许可协议：署名-相同方式共享 4.0 国际 (CC BY-SA 4.0)

<h2 id="中国娱乐圈关系挖掘">中国娱乐圈关系挖掘</h2>
<blockquote>
文章最后有系统的传送门
</blockquote>
<hr>
<blockquote>
作者博客的传送门 ，小弟在校学生，初来乍到，第一次发博客，请喷。我希望能不断的学习，然后提高自己。
</blockquote>
<h2 id="题目描述">题目描述</h2>
中国的娱乐圈关系网十分庞大，存在着紧密的联系。在学习完田老师的复杂网络课程之后，我们中国的中国娱乐圈关系挖掘产生了浓厚的兴趣。在这次的大作业中，我们将会探索中国娱乐圈中任何个人之间的关系，测试六度分割理论，并进行更深入的尝试！
<h2 id="数据采集">数据采集</h2>
<h2 id="关系采集">关系采集</h2>
本实验使用的数据主要是在互动百科和百度百科中的物关系一栏进行提取。如下图所示，分别是“周杰伦”在互动百科和百度百科中的人物关系。并进行单的去重和筛选，便于进行下面的实验。
<img src="https://ld246.com/images/img-loading.svg" alt="c59eaed11221491399f8d27d91cec60-1.png" data-src="https://b3logfile.com/file/2017/2/c59eaed11221491399f8d27ad9cec60-1.png?imageView2/2/interlace/1/format/jpg">
<img src="https://ld246.com/images/img-loading.svg" alt="d4d3c846644f4565897655a34eca017-2.png" data-src="https://b3logfile.com/file/2017/2/d4d3c846644f4565897655a234ca017-2.png?imageView2/2/interlace/1/format/jpg">
我们在数据采集结束之后，将这些关系存到了 neo4j 中，这里我们对 neo4j 进行简单的介绍。Neo4j 是一个图形数据库，这也就意味着它的数据并非保存在表或集合中，而是保存为节点以及节点之的关系。在 Neo4j 中，节点以及关系都能够包含保存值的属性，此外：

可以为节点设置零或多个标签（例如 Author 或 Book）

每个关系都对应一种类型（例如 WROTE 或 FRIEND_OF）

关系总是从一个节点指向另一个节点（但可以在不考虑指向性的情况下进行查询）

最终，我们拥有 193 种类型的关系，一共 3830 条关系，数据量不是很大，所以我们的大作业仅是做一个实验性的工作。
<h2 id="图像采集">图像采集</h2>
主要在百度图片中进行爬取，鉴于百度图片的防爬虫机制，而且我们的人物总数不多，采用了模浏览器的方法，强行进行图片的抽取。最终我们获取了 2205 张图片，即代表我们有这么多个实体（星）数量，如下图所示。
<img src="https://ld246.com/images/img-loading.svg" alt="066e3321e1564279b5d0b7f429fbdcd-3.png" data-src="https://b3logfile.com/file/2017/2/066e3321e1564279b5d0b7f942fbdcd-3.png?imageView2/2/interlace/1/format/jpg">
这里主要讲述后台使用的算法。
<h2 id="最短路径查找">最短路径查找</h2>
最短路径的算法我们使用了三种主流的算法，在结果的展示方面，结果可能会有不同，因为采用算法，会导致路径查找的策略不同，进而结果也不同。这里我们仅仅展示一个 Floyd 算法搜索的结果更多的结果可以去系统里进行亲手尝试。如下图，展示的是“鹿晗”和“张国立”的关系。
<img src="https://ld246.com/images/img-loading.svg" alt="658a2abedd3e476b829109

bb80b71b8-4.png" data-src="https://b3logfile.com/file/2017/2/658a2abedd3e476b8291090b80b71b8-4.png?imageView2/2/interlace/1/format/jpg">
<h3 id="Floyd算法">Floyd 算法</h3>
Floyd-Warshall 算法（Floyd-Warshall algorithm）是解决任意两点间的最短路径的一种算法可以正确处理有向图或负权的最短路径问题，同时也被用于计算有向图的传递闭包。Floyd-Warshall 算法的时间复杂度为 O(N3)，空间复杂度为 O(N2)。
Floyd 算法是一个经典的动态规划算法。用通俗的语言来描述的话，首先我们的目标是寻找从点 i 到点 j 的最短路径。从动态规划的角度看问题，我们需要为这个目标重新做一个诠释（这个诠释正是态规划最富创造力的精华所在）
从任意节点 i 到任意节点 j 的最短路径不外乎 2 种可能，1 是直接从 i 到 j，2 是从 i 经过若干个点 k 到 j。所以，我们假设 Dis(i,j)为节点 u 到节点 v 的最短路径的距离，对于每一个节点 k，我们检 Dis(i,k) + Dis(k,j) < Dis(i,j)是否成立，如果成立，证明从 i 到 k 再到 j 的路径比 i 直接到 j 的路径，我们便设置 Dis(i,j)
= Dis(i,k) + Dis(k,j)，这样一来，当我们遍历完所有节点 k，Dis(i,j)中记录的便是 i 到 j 的最短路径的离。
<h3 id="Dijkstra算法">Dijkstra 算法</h3>
Dijkstra(迪杰斯特拉)算法是典型的单源最短路径算法，用于计算一个节点到其他所有节点的最路径。主要特点是以起始点为中心向外层层扩展，直到扩展到终点为止。Dijkstra 算法是很有代表性最短路径算法，在很多专业课程中都作为基本内容有详细的介绍，如数据结构，图论，运筹学等等。意该算法要求图中不存在负权边。
算法思想：设 G=(V,E)是一个带权有向图，把图中顶点集合 V 分成两组，第一组为已求出最短路的顶点集合（用 S 表示，初始时 S 中只有一个源点，以后每求得一条最短路径，就将加入到集合 S 中直到全部顶点都加入到 S 中，算法就结束了），第二组为其余未确定最短路径的顶点集合（用 U 表），按最短路径长度的递增次序依次把第二组的顶点加入 S 中。在加入的过程中，总保持从源点 v 到 中各顶点的最短路径长度不大于从源点 v 到 U 中任何顶点的最短路径长度。此外，每个顶点对应一距离，S 中的顶点的距离就是从 v 到此顶点的最短路径长度，U 中的顶点的距离，是从 v 到此顶点只括 S 中的顶点为中间顶点的当前最短路径长度。
<h3 id="Bellman-Ford算法">Bellman-Ford 算法</h3>
Bellman-Ford 算法是一种用于计算带权有向图中单源最短路径（SSSP：Single-Source Shortes Path）的算法。该算法由 Richard Bellman 和 Lester Ford 分别发表于 1958 年和 1956 年，而实上 Edward F. Moore 也在 1957 年发布了相同的算法，因此，此算法也常被称为 Bellman-Ford-More 算法。
Bellman-Ford 算法和 Dijkstra 算法同为解决单源最短路径的算法。对于带权有向图 G = (V, E) Dijkstra 算法要求图 G 中边的权值均为非负，而 Bellman-Ford 算法能适应一般的情况（即存在负权的情况）。一个实现的很好的 Dijkstra 算法比 Bellman-Ford 算法的运行时间要低。
Bellman-Ford 算法采用动态规划（Dynamic Programming）进行设计，实现的时间复杂度为 O V*E)，其中 V 为顶点数量，E 为边的数量。Dijkstra 算法采用贪心算法（Greedy Algorithm）范式行设计，普通实现的时间复杂度为 O(V2)，若基于 Fibonacci heap 的最小优先队列实现版本则时间杂度为 O(E + VlogV)。
<h2 id="个人关系网">个人关系网</h2>
输入一个明星，挖掘出与其相关的明星。主要就是数据库的查找，为此，我们专门研究了 neo4j 的数据库的查询等方面的知识。
如图所示，我们展示了“林俊杰”的个人关系网。
<img src="https://ld246.com/images/img-loading.svg" alt="db54a5fb83234dc8bf350932e90f9f1-5.png" data-src="https://b3logfile.com/file/2017/2/db54a5fb83234dc8bf35093d2e0f9f1-5.png?imageView2/2/interlace/1/format/jpg">
<h2 id="日志记录与挖掘">日志记录与挖掘</h2>
为了更好地分析用户数据，我们特意加入了日志系统，在后台把所有的日志进行了收集，并在前进行了可视化展示。
<img src="https://ld246.com/images/img-loading.svg" alt="0d6d93650db74ef6b1d4eefc990ef9a-6.png" data-src="https://b3logfile.com/file/2017/2/0d6d93650db74ef6b1d4eef6c90ef9a-6.png?imageView2/2/interlace/1/format/jpg">
在日志系统中，我们记录了搜索的关键字、使用的算法、搜索用时、日期、ip 等常用信息，并对些简单的数据进行了挖掘和分析，比如，我们统计了搜索最多的关系，统计了最热的明星是谁？当然

，目前大多是我们的测试数据。很高兴的是，今天我们迎来了第一位真实用户，并尝试了大量的搜索显然，他对我们的系统很是喜欢！
上图展示的是搜索最多的词，和最热的明星。其中大多数都是我自己测试使用的，并不能代表使我们系统的用户的观点和情感倾向。
<img src="https://ld246.com/images/img-loading.svg" alt="047ebd6319e24eb2adf75ee309eaf55-7.png" data-src="https://b3logfile.com/file/2017/2/047ebd6319e24eb2adf75eea30eaf55-7.png?imageView2/2/interlace/1/format/jpg">
<h2 id="后台">后台</h2>
后台主要使用 java 进行开发，这里主要讲采用的技术框架。
<h3 id="Jersey-2-x">Jersey 2.x</h3>
我们的工程主要是使用这个组件进行 restful
API 的开发，之后的开发中，我将会完全转向 spring 的家族中。
Jersey 是 JAX-RS 的参考实现，它包括三个主要部分：核心 server（Core Server）：通过提供 JR 311 中标准化的凝视和 API 标准化，您能够用直观的方式开发 RESTful Web 服务。核心 client（Core Client）：Jersey
client API 帮助您与 REST 服务轻松通信。集成（Integration）：Jersey 还提供能够轻松集成 Sprin、Guice、Apache Abdera 的库。
<h3 id="Swagger">Swagger</h3>
Swagger 是一款 RESTFUL 接口的文档在线自动生成 + 功能测试功能软件。本文简单介绍了在目中集成 swagger 的方法和一些常见问题。如果想深入分析项目源码，了解更多内容，见参考资料。/p>
Swagger 是一个规范和完整的框架，用于生成、描述、调用和可视化 RESTful 风格的 Web 服。总体目标是使客户端和文件系统作为服务器以同样的速度来更新。文件的方法，参数和模型紧密集到服务器端的代码，允许 API 来始终保持同步。Swagger 让部署管理和使用功能强大的 API 从未如简单。
<h2 id="前台">前台</h2>
首先，我先挂上几张图对我们的前台进行展示，尽管之前已经展示了部分细节。
<img src="https://ld246.com/images/img-loading.svg" alt="b5429c8cc87446b2b17dd700efbffbd-8.png" data-src="https://b3logfile.com/file/2017/2/b5429c8cc87446b2b17dd7090ebffbd-8.png?imageView2/2/interlace/1/format/jpg">
<img src="https://ld246.com/images/img-loading.svg" alt="6491d2741ab54ae598ce1ed02d80b8d-9.png" data-src="https://b3logfile.com/file/2017/2/6491d2741ab54ae598ce1edc0d80b8d-9.png?imageView2/2/interlace/1/format/jpg">
如上图所示，第一个页面展示的是关系搜索结果，并将每一次的结果往下推，新的结果展示在最面。第二个页面是我们的日志分析系统，之后可以日志挖掘做的再完美一点！再实用和新颖一点！</>
<h3 id="AngularJS">AngularJS</h3>
前台的主要 js 框架是 angular 的框架，使用简单，不过 1.x 版本有点老了，虽然大多数的网站是在使用 1.x，之后会考虑使用国产的 vue.js，该框架在使用方面会更加便捷！
AngularJS 是一个 JavaScript 框架。它可通过 标签添加到 HTML 页面。AngularJS 通过 指令展了 HTML，且通过 表达式 绑定数据到 HTML。
<h3 id="D3-js">D3.js</h3>
我们的主要的绘图工作都是它完成的，在绘图的专业性方面，首推它。
D3.js（D3 或 Data-Driven Documents）是一个用动态图形显示数据的 JavaScript 库，一个数可视化的工具。兼容 W3C 标准，并且利用广泛实现的 SVG，JavaScript，和 CSS 标准。它是早期的rotovis 框架的继承者。与其他的类库相比，D3 对视图结果有很大的可控性。D3 是 2011 年面世的同年的 8 月发布了 2.0.0 版。到 2012 年 12 月，D3 已被更新到了 3.0.0 版。
<h3 id="Bootstrap">Bootstrap</h3>
在系统的样式方面，使用了大量的线程的组件，这样可以极大的方便我们进行 css 开发，并帮助们在视觉设计等方面欠缺的工程师在前台上进行更好的开发。
Bootstrap，来自 Twitter，是目前最受欢迎的前端框架。Bootstrap 是基于 HTML、CSS、JAVSCRIPT 的，它简洁灵活，使得 Web 开发更加快捷。
<h2 id="小组成员以及分工">小组成员以及分工</h2>
<h2 id="石磊">石磊</h2>

总工程师

数据处理

可视化

后台框架制定

实现 2 个算法

<h2 id="郭朝彤">郭朝彤</h2>

总助理

数据处理

图片爬虫

实现一个路径搜索算法

翻译论文

<h2 id="总结">总结</h2>
这次的大作业自己的收获很大，将自己在课程学到的东西和生产实际进行了结合，是我们能够在码中体会到复杂网络的快乐！我们很喜欢田老师的课程，希望以后大家都能够选这门课，将复杂网络乐趣传递下去！
<h2 id="在线预览">在线预览</h2>
传送门
该服务器是学校的，离校之后不保证正常访问。若想了解，请 email！
服务器由于物理原因宕机不可修复，数据没了。重新发布的。