1 背景研究

随着现代信息技术的不断发展, 全球绝大部分国家已进入了互联网时代。如今大多数年轻人主要求职方式是借助网络,企业进行招聘的主要方式也逐渐基于网络。因此，目前在网络上已经诞生了各式各样的综合型招聘网站，例如Indeed，Liepin。企业在这些网站上发布岗位信息，应聘者根据需要进行选择。然而在网络招聘的过程中也存在弊端性。在Indeed网站上对于每一项所发布的职位信息，还会有一个发布时间显示出来。如果发布时间距离现在的时间越长就说明了两个问题，其一是这个岗位人才确实比较稀缺，另一个原因就是企业不喜欢用网络招聘。通过对于多个综合性招聘网站的实际观察，我发现它们都缺乏对于岗位信息数据的综合展示区域。目前随着综合型招聘网站的数据不断积累，已经有许多人想要从中挖掘有用的信息来进行分析。D.Smith从招聘网站收集有关程序开发人员的招聘信息, 并利用关键词索引技术研究企业对求职者编程能力的需求趋势, 为高校计算机专业课程设置提供重要参考。JIA通过检索中国拉勾网中数据分析师这一岗位信息来获得岗位情况。他的分析不仅岗位单一而且仅仅只分析了一个招聘网站，这样会导致数据不够全面。除此之外，国内外相关研究大多采用传统的抽样方式，对招聘网站数据进行人工统计整理, 数据量小、效率低, 不适合大数据环境下快速且智能化的数据采集与分析。

With the continuous development of modern information technology, most countries in the world have entered the Internet era. Nowadays, most young people mainly apply for jobs through the Internet, and the main way for enterprises to recruit is gradually based on the Internet. Therefore, a variety of comprehensive recruitment websites have emerged on the Internet, such as India and Liepin. The enterprise publishes position information on these websites, and the candidates choose according to their needs. However, there are also disadvantages in the process of online recruitment. There will also be a release time for each published position information on the company's website. The longer the publishing time is from now, the two problems will be explained. One is that talents in this position are indeed scarce, and the other is that enterprises do not like to use Internet recruitment. Through the actual observation of several comprehensive recruitment websites, I found that they lack comprehensive display area for position information data. At present, with the continuous accumulation of comprehensive recruitment website data, there are many people want to mine useful information for analysis. D. Smith collects recruitment information about program developers from the recruitment website, and uses keyword index technology to study the demand trend of enterprise for the programming ability of job seekers, so as to provide important reference for the curriculum setting of computer major in Colleges and universities. Jia obtains the position information by retrieving the position information of data analyst in China's Lagoon network. His analysis is not only a single job but only a recruitment website, which will lead to incomplete data. In addition, most of the relevant researches at home and abroad adopt the traditional sampling method to sort out the recruitment website data by artificial statistics. The data volume is small and the efficiency is low, which is not suitable for the rapid and intelligent data collection and analysis in the big data environment.

Web信息挖掘

在通过阅读了一部分文献之后，我发现在对于Web数据信息获取的方法中，人们普遍运用Python进行数据挖掘，但是也有人会选择利用JAVA来开发爬虫程序。接下来我将会分别讨论利用Python和JAVA进行Web爬虫的开发。

XPath是一门在XML文档中查找信息的语言,用于在XML文档中通过元素和属性进行导航。利用XPath可以方便地在HTML文档中定位感兴趣的节点。lxml库是Python的第三方库,它支持标准的XPath规范。可对XML文档中标签进行导航和提取标签内容的语言。在XPath中,有5个基本类型的节点:文档(根)节点、类别、文本、属性、元素。在对各个节点寻找时为了区分是找的根节点还是类别,XPath语言规定需要在名称前加特殊的符号。使用前需从lxml库中导入etree包。

代码如下所示:

Beautiful soup没有自己的解析器。其在构建一个对象的时传入一个参数,以指定解析器。它支持的众多的解析器中,可以使用自己喜欢的解析器.