第一周进展报告

计 92 丘骏鹏 2009011282

2013-03-05 Tue

1 每天进展

- 周二：安装实验室电脑，配置基本编程环境
- 周三周四：阅读史兴的论文。有以下几个问题：

  1. 关于 data records 的检测方法（基于桶）好像有鲁棒性不够的问题，若出现噪声标签会产生问题
  2. ATree 的构造方法还有一些细节方面感觉论文未详细论述

  除此之外，对于论文中提到的一些技术，如 Hadoop，在网上自学了关于 HFS、HBase 等知识以及简单的 MapReduce 程序编写。其他的比较细的东西包括 Google Protocol Buffers 以及一些 HTML 清洗库（HTML Tidy，lxml.html 模块等）。

- 周五：从图书馆和网上找了有关 dom tree clustering 的相关资料。主要阅读了《XML 挖掘: 聚类、分类与信息提取》一书，了解了一些计算两个 XML 文档相似性和进行聚类的方法。

2 对工作的理解

主要工作分为两部分，一部分是从网页中分类出所需要的网页，一部分是对网页进行聚类，然后生成模板进行提取。

- 第一部分工作我认为主要可以通过 url 来进行判断，一般目录页和内容页的 url 有较大差别，这种方法会非常有效率。如果这种方法不行，比如已有的网页的 url 全是随机生成的，是可以通过一些简单的特征，比如文本元素的内容长度进行筛选。这部分复杂度应该不能做得很高。
- 第二部分工作是此次毕设的主要部分。主要的问题应该是“如何对 Dom Tree 进行聚类，并从每个类别中提取出模板”（有无理解错误?）。这部分还在阅读相关的书籍和论文，打算结合这周阅读论文的成果在这周的报告中给出。

3 问题

- 目前有的资源有哪些？我看了上次您发给我的软院那个同学的东西，好像只是调研了 Java 的各个库如何抓取用了 ajax 的网页，应该只是和写网络爬虫那一方面相关的。目前我们有没有相关的网页，是需要自己抓取，还是已经存好了？如果已经有的话，如何获取使用？如果需要自己抓取数据或者预处理起来还有很多繁琐的细节，我觉得也可以和读论文同步展开。