第四周工作报告

计 92 丘骏鹏 2009011282

2013-03-22 Fri

1 工作内容

- A Short Survey of Document Structure Similarity Algorithms。这篇文章总结了已有的几种计算 HTML 文档结构相似性的方法，提出了 Path Shingles 的计算方法，文章中还提到了几种计算方法的若干实验结果，也可以当作本实验的参考。
- Clustering Template Based Web Documents。这篇论文讨论一种优化的树的编辑距离计算方法和几个基于 Path 和 Tag 的相似度计算方法，同时还有将这些计算方法和几种聚类方法结合以后的实验效果，论文中的一些评价方法可以借鉴。
- Detecting structural similarities between XML document。这篇文章只是简单的浏览，主要讲的是如何用傅里叶变换进行相似度计算。结合其他引用了这篇文章的论文来看，此方法比较复杂，效果却不一定很好，因此暂时不考虑使用。
- 其他的大部分时间在准备、修改和完善开题报告

2 下周计划

主要是开始写代码，将数据的输入方式和预处理做好，同时初步实现网页过滤的功能。可能在实现过程中还要继续看一些论文，研究其中的细节。根据代码实现的进度再灵活调整工程进度。