第二周报告

丘骏鹏

2013-03-08 Fri

目录

1 本周进展

本周主要工作在阅读相关资料和论文。
中文部分：

- 《XML 挖掘：聚类、分类与信息提取》：这本书主要是一些综述，讲解了一些基本概念和方法。
  - 关于 XML 等一系列的概念：DTD，Schema，XSL，XPath，XQuery 等
  - 基本的计算 XML 文档相似度的方法：简单的相似度度量，路径匹配，编辑距离，向量空间模型等。不过主要是方法综述，未讲解实现细节。
  - 书中提出的方法：类似路径匹配，加入 WordNet 进行语义消歧，然后计算出文档之间的相似度
- 《基于局部标签树匹配的改进网页聚类算法》：这篇论文主要是将 DOM Tree 的前几层转化为字符串，然后利用字符串的编辑距离衡量 DOM Tree 的相似度。具体方法为：先将每层的标签拼接成字符串，计算相似度，然后不同的层级给予不同的权重，相加得到总的相似度。方法比较简单。
- 《基于网页聚类的 web 信息自动抽取》：这篇论文主要基于 DOM Tree 的编辑距离，然后计算出平均绝对误差的列相似度，再利用 CURE 算法进行聚类。CURE 聚类算法的论文我看的是 http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.1599&rep=rep1&type=pdf

英文部分：

- UIC Bing Liu 的 CS511 课程课件，讲的是结构化数据抽取中有监督和无监督的 wrapper 生成算法。里面包括了树相似度算法的介绍，有 Tree Edit Distance，Simple Tree Matching （STM），Center Star method 和 Particial Tree Alignment。
- Yanhong Zhai and Bing Liu. Web Data Extraction Based on Partial Tree Alignment. 选择性看了关于 Partial Tree Alignment 的部分。

以下两篇是史兴的论文中的参考文献。

- A. Arasu and H. Garcia-Molina. Extracting structured data from web pages. 主要讲解的是通过构造 Equivalence Class，进行模板的生成。
- A. Carlson and C. Schafer. Bootstrapping information extraction from semi-structured web pages. 首先用 partial tree alignment 进行对齐，然后利用一些语义特征，对这些可能的数据域和 schema 中的某个 column 的相似度进行打分，训练模型采用 bootstrapping。

在读的有一篇：TEXT: Automatic Template Extraction from Heterogeneous Web Pages，文中提出一种效率较高的计算相似度的方法，具体细节部分还在看。您给的那篇论文目前还没有看。

2 问题

Google 开源的页面抽取工具我没找到，能否给个链接？

3 接下来的工作

- 这周刚开始看论文，感觉方向有点杂，接下来要做好总结和细化工作。
- 看一下相关的工具，包括有些论文里提到的工具，看下能否使用以及如何使用，避免重复造轮子。