毕设工作

丘骏鹏

2013-03-13 Wed

1 工作综述

我对主要工作的理解是：从海量的网页数据中，利用大数据的冗余性发现网页模板，抽取结构化的信息。主要的步骤：

- 首先需要将目录页和详细页分开，可以根据 url 特征和简单的文本特征来区别。这部分应该要比较简单迅速。
- 工作的重点在于如何对海量的网页进行聚类和模板抽取。这部分工作应该要充分考虑数据的冗余性，利用这些冗余信息来发现不同网页之间的共同点，从而自动生成网页抽取的模板。同时考虑到数据量很大，算法不能太复杂。

2 目前的工作进展

前两周的工作报告放在了 report_week1-2.zip 中。
这周到目前截止的工作：

- 阅读了 TEXT: Automatic Template Extraction from Heterogeneous Web Pages。这篇论文将 DOM Tree 用路径进行表示，用 MDL(Minimun Description Length Principle) 作为标准进行聚类。由于直接计算 MDL 复杂度很高，该论文提出一种扩展的 MinHash 算法，用于近似估计 MDL，提高算法效率，
- 阅读了黄老师给的 Webpage Understanding: Beyond Page-Level Search。里面主要介绍的是如何将 webpage understanding 拆分成几个子任务。其中提到了一种分割 HTML 网页的 VIPS(VIsion-basd Page Segmentation) 方法，即将原网页按照视觉区域的分块解析成一个 vision tree 来进行表示。
- VIPS: a Vision-based Page Segmentation Algorithm（在读）。即上面提到的 VIPS 算法具体实现的论文。我初看的结果是好像算法比较复杂，不确定是否适合应用到大数据上面。
- http://code.google.com/p/cx-extractor/ 上的工具和文档。该工具主要就利用了 HTML 文档行块的分布来提取正文，方法和实现都很简单。

3 我对工作的理解

工作的重点应该是网页的聚类和模板抽取。

对于如何进行聚类。我觉得聚类方法可以采用现有的那些聚类算法，关键是如何有效地计算两个文档的相似度。HTML 可以解析成 DOM Tree，可以直接利用这个树表示计算编辑距离来衡量相似度，也可以将树通过某种遍历方法转化为其他数据结构再进行计算，或者是用路径集合表示，也可以按照 VIPS 将网页划分成 vision-tree 后再做计算。我现在还没有想到一个很好的计算方法。

对于模板如何抽取。如果聚类的时候利用了 DOM Tree 的结构特征，比如标签，路径，模板就自动可以表示出来了。如果利用纯内容特征（应该不会这样做），可能会涉及到如何进行正则表达式推导问题。