第三周报告

丘骏鹏

2013-03-15 Fri

1 这周的工作

前半周的工作已经在上一个报告中给出了，这里将上一份报告的对应部分复制过来，另外汇报一下下半周的工作。

- 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。
- 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。
- VIPS: a Vision-based Page Segmentation Algorithm. 这篇论文讲解的是 VIPS 的实现。方法比较复杂，由于文章比较老 (2003)，网页的设计有很大的变化，论文中列出的一些启发规则可能要有一定的修改。此外，我找了一下，作者没有公开实现的代码。
- 阅读了中文文献《基于视觉的 Web 页面分块算法的改进与实现》，是一个对 VIPS 算法的改进。主要关注的 HTML 中的 <table> 标签，将 DOM Tree 变成只有 <table> 标签的树（实际上对于 是一样的，现在网页大部分已经采用 DIV+CSS 的布局方法），同时不需要用规则判断节点是否可分，直接提取每一层的 <table> 标签子节点，后面还有对各个语义块进行合并的过程。

  我原来的想法是有没有可能在 vision-tree 的层次上做聚类，相同模板的网页视觉布局大概也会相同。但是现在网页设计有一些变化，布局信息更多放在 CSS 中，因此如果网页设计者将 HTML 的布局和内容分开得比较好的话，HTML 源代码中大部分会是内容，布局信息会很少，而这个方法主要是依靠 tag 里的一些视觉相关的属性值进行判断，加上这个系统比较复杂，我没找到现成的实现（好像郝老师说有类似实现），可能可行性不高。

2 我的想法

从目前的调研结果来看，我觉得 TEXT: Automatic Template Extraction from Heterogeneous Web Pages 这篇论文和我要做的工作比较接近，主要关注的也是从异构的网页中抽取模板。他的基本特征是基于 HTML 的 DOM Tree 的路径集合，改进的地方在于提出了一个近似算法（Extended MinHash）对计算进行加速。关于 Extended MinHash 算法的很多细节我当时没有全部看懂，我打算再认真看看论文的那一部分。

现阶段我的想法是基于 vision-tree 的方法可能不太可行，还是基于原始的 DOM Tree 进行聚类。模板的表示基于路径会好一些，但聚类时候用的特征不一定局限于路径，可以结合标签加上文本内容特征。另外，我之前看的《XML 挖掘：聚类、分类和信息提取》书中的利用 WordNet 进行路径的词义消歧可能也会有一些帮助。

我搜索了一些其他在直接在 DOM Tree 上利用路径、token 进行聚类的论文，接下来几天认真调研直接在 DOM Tree 上做的方法。