当前组件对ＨＴＭＬ５站点的适用性

测试报告

黎强（[liqiang@software.ict.ac.cn](mailto:liqiang@software.ict.ac.cn)）

# 测试目的

测试当前已有的四个组件：Prophet、Constor、Indilator、MASK对HTML5页面的抽取适应能力。

# 测试方法

**测试环境：**WDE-IE-新闻版（http://10.61.1.44/content/news/login.html）

**测试方法：**利用WDE-IE-新闻版，对找到的各个html5站点进行采集，然后分别用各个组件（Prophet、Constor、MASK、Indilator）对各个站点进行配置与抽取，然后统计如下信息：

* 某一站点采集到的文章数
* 各个字段抽取成功的文章数
* 各个字段在所有站点上总的抽取成功率

# 测试站点

|  |  |  |  |
| --- | --- | --- | --- |
| 网站名称 | 网站入口URL | Blog正则 | 类型 |
| forgottenpresidents | <http://forgottenpresidents.com/> | [http://forgottenpresidents\.com/.+?\.html](http://forgottenpresidents\\.com/.+?\\.html) | Blog |
| iconeden | <http://www.iconeden.com/blog/> | [http://www\.iconeden\.com/blog/\d{4}/\d{2}/\d{2}/.+?/](http://www\.iconeden\.com/blog/\d%7b4%7d/\d%7b2%7d/\d%7b2%7d/.+?/) | Blog |
| ourswisslife | <http://www.ourswisslife.com/> | [http://www\.ourswisslife\.com/blog/\d{4}/\d{2}/.+?](http://www\\.ourswisslife\\.com/blog/\\d%7b4%7d/\\d%7b2%7d/.+?) | Blog |
| girlwithacamera | <http://girlwithacamera.co.uk/> | [http://girlwithacamera\.co\.uk/[^/]+?/](http://girlwithacamera\\.co\\.uk/[^/]+?/)$ | Blog |
| webdesignwall | <http://webdesignerwall.com/> | [http://webdesignerwall\.com/tutorials/.+?](http://webdesignerwall\.com/trends/.+?) | Blog |

# 测试结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 网站名称 | Prophet | | | |
| 正文  (召回，准确) | 标题  (召回，准确) | 发布时间  (召回，准确) | 作者  (召回，准确) |
| forgottenpresidents |  |  |  |  |
| iconeden | 7/7 7/7 | 7/7 7/7 | 7/7 7/7 | 0/7 0/0 |
| ourswisslife | 5/5 3/5 | 5/5 5/5 | 5/5 5/5 | 0/0 0/0 |
| girlwithacamera | 8/8 8/8 | 8/8 8/8 | 0/8 0/0 | 0/0 0/0 |
| webdesignwall | 8/8 8/8 | 8/8 8/8 | 0/8 0/0 | 0/0 0/0 |
| ALL | 100% 92.9% | 100% 100% | 42.9% 100% | 0% 100% |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 网站名称 | Constor | | | |
| 正文  (召回，准确) | 标题  (召回，准确) | 发布时间  (召回，准备) | 作者  (召回，准确) |
| forgottenpresidents | 4/4 4/4 | 4/4 0/4 | 0/0 0/0 | 0/0 0/0 |
| iconeden | 7/7 7/7 | 7/7 7/7 | 7/7 7/7 | 0/7 0/0 |
| ourswisslife | 5/5 5/5 | 5/5 5/5 | 5/5 5/5 | 0/0 0/0 |
| girlwithacamera | 8/8 1/8 | 5/8 5/8 | 8/8 8/8 | 0/0 0/0 |
| webdesignwall | 8/8 7/8 | 8/8 8/8 | 8/8 8/8 | 0/0 0/0 |
| ALL | 100% 75% | 100% 87.5% | 100% 100% | 0% 100% |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 网站名称 | MASK | | | |
| 正文  (召回，准确) | 标题  (召回，准确) | 发布时间  (召回，准备) | 作者  (召回，准确) |
| forgottenpresidents | 4/4 4/4 | 4/4 4/4 | 0/0 0/0 | 0/0 0/0 |
| iconeden | 7/7 7/7 | 7/7 7/7 | 7/7 7/7 | 7/7 7/7 |
| ourswisslife | 5/5 5/5 | 5/5 5/5 | 5/5 5/5 | 0/0 0/0 |
| girlwithacamera | 8/8 8/8 | 8/8 8/8 | 8/8 8/8 | 0/0 0/0 |
| webdesignwall | 8/8 8/8 | 8/8 8/8 | 8/8 8/8 | 0/0 0/0 |
| ALL | 100% 100% | 100% 100% | 100% 100% | 100% 100% |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 网站名称 | Indilator | | | |
| 正文  (召回，准确) | 标题  (召回，准确) | 发布时间  (召回，准备) | 作者  (召回，准确) |
| forgottenpresidents | 4/4 4/4 | 4/4 4/4 | 0/0 0/0 | 0/0 0/0 |
| iconeden | 7/7 7/7 | 7/7 7/7 | 7/7 7/7 | 7/7 7/7 |
| ourswisslife | 5/5 5/5 | 5/5 5/5 | 5/5 5/5 | 0/0 0/0 |
| girlwithacamera | 8/8 8/8 | 8/8 8/8 | 8/8 8/8 | 0/0 0/0 |
| webdesignwall | 8/8 8/8 | 8/8 8/8 | 8/8 8/8 | 0/0 0/0 |
| ALL | 100% 100% | 100% 100% | 100% 100% | 100% 100% |

# 分析与结论

**分析：**

首先，本次测试中，所找到的站点全部为个人blog。主要原因有两点：

1. html5的特点主要在多媒体上，而我们的组件针对的均为文本内容。所以大量的html5站点无法参与测试
2. 对于以文本为主要内容的新闻、bbs站点。因为html5与html4相比，虽然可以有更好的结构，但是因为html4基本可以满足需求，且新闻站点一般为大公司或者机构所有，内容非常多，且html4可以满足需求，而升级为html5带来的效果不明显，但成本高，所以，几乎没有完整的html5新闻站点；而对于bbs，大部分采用discuz等成熟的框架，而这些框架对html5的需求并不明显，且同样升级成本高，因而也几乎没有完整采用html5的bbs站点。因此，只留下了个人博客，既可以满足我们的文字处理要求，又因为个人开发者对新技术的尝新，以及有些博客框架升级为html5仅需要升级主题，而这些主题又是可以有大量的个人开发者设计和开发的，所以相对站点较多。

然后，从测试结果中可以看出：

1. 总体上，各个组件对html5网页的抽取并没有出现明显的不适应。与html4网页的抽取水平基本相同。同时，从本人的标注过程来看，也并没有碰到需要专门针对html5做特殊处理。
2. 分组件来看，自动生成模板的组件准确率与召回率都在一定程度上比手动模板标注低。同时，从两个手动标注模板的组件各个属性的召回率跟准备率均为100%可以看出，只要模板标注合适，抽取成功率还是非常高的。这主要是html5并没有对html的结构做根本性的改变，而仅仅是对节点做了一些改进，所以，对于基于dom树结构的抽取方法基本没有影响。

**结论：**

HTML5普及还需要一定的时间，主要应用范围为多媒体等，不在我们抽取的考虑范围内，而对于blog等，现有组件依然适用。