基于具体需求的URL辅助法实际应用深入研究

# 一、研究背景及思路

**研究目的：**通过提取URL的特征信息对网页作出准确的分类。

**研究现状：**传统对网页分类的方法主要基于两类特征。第一类是页面特征；第二类是邻居特征，页面特征里面又包括正文、锚文本、标题等等；但是这却忽略了一个重要信息，页面自身的链接信息，互联网上任何一个页面都是有一个唯一的标识，那就是它的链接。而且传统分类方法分类速度慢、需要下载全部网页、有些分类信息隐藏在图片中等。通过搜索论文发现传统上基于URL特征网站分类方法，希望通过一个统一的方法解决互联网上网站所有网址的分类问题，并且也提出了很多有效的方法，但是都没有一种方法准确率会特别高。

**本文方法特点：**基于URL对网页分类的优势很多。URL容易被人所记忆，良好的网站设计技术和编码通过有用的词汇将网页的信息大量的说明，主题隐含在URL中也是经常可见。一个URL的时序依赖性、大小写、特征长度、以及URI组件特征，对网页分类都是有用的。本文提出一种基于需求的URL辅助法，该方法不企图一次性解决互联网所有分类的问题，通过严格的限制从而提高分类的精度。

**本文方法缺点：**严重依赖需求，严重依赖网站设计。

**特征提取具体实施步骤：**

第一步：确定客户需求，明确客户需要检测的网站。

第二步：确定各个网站的导航网址，判断网站是否适合通过URL来进行分类。

第三步：指定网址地图分类块抽取规则，提取各个分类对应的网址。

第四步：保存分类信息。

**链接分类具体实施步骤：**

第一步：判断链接是否是在需求范围内，如果是执行第二步，如果不是执行第三步。

第二步：判断链接属于哪个分类，层层判断，一直到无法判断那一层停止。

第三步：提示链接不在需求范围内，提取关键词，根据关键词匹配算法，为此链接分配一个类，如果一个关键词都没能匹配，随机分配一个类，并且提示错误信息。

**希望取得的成果：**

通过这种半自动化的方式，最终能够对需求范围内的网页完成比较高的主题分类。

# 二、具体实现

实现工程命名为：urlparser 。

2.1工程简介

说明：urlparser 目的是通过整合所有需求网址网站导航的分类信息。

例如新浪网：http://www.sina.com.cn/ 它的导航页是：http://news.sina.com.cn/guide/

提取导航页所有分类锚文本以及链接，作为样本数据，下面会提到具体的样本规则化的过程。

这样对于新的，通过计算距离哪个样本最近来确定它的分类。

本工程所有的包以 eshore.cn.it作为前缀，代表广州亿迅IT部门。主要分为四块：

data 用于处理数据，包括数据读写过滤。

DataFilter.java 根据resource目录中的filterwords.txt过滤数据

DataLoader.java 加载数据，包括初始网站数据，提取之后的样本数据加载

DataWriter.java 数据持久化，包括提取到的样本保存到data目录下的keywords.txt

KeyWord.java 样本数据对应类

SiteMap.java 网站对应类（包括网站主页网址，网站导航网址，网页分类块选择字符串）

main 程序入口

ParseStart.java 对于新的网站，提取样本数据

parser 关键逻辑代码

LabelRecognizer.java 链接类别识别类

UrlParser.java 样本数据提取类

## 2.2 工程使用

### 2.2.1 添加样本数据：

在sitemaps.txt中按照如下格式添加网站：

```

id,网站中文名,主页网址,导航页网址,分类块选择

1,新浪网,http://www.sina.com.cn/,http://news.sina.com.cn/guide/,div#tab01 div.clearfix

```

必须保留第一行，每一列通过逗号隔开，每一列的意义如第一行所示。其中分类块选择字符串请查考Jsoup select 使用，具体网址如下：

http://jsoup.org/

主程序入口：eshore.cn.it.main.ParseStart.java

程序会自动保存样本数据到data目录下的keywords.txt,具体格式如下：

```

id,pid,url,host,keywords,label

1,-1,http://www.sina.com.cn/,www.sina.com.cn,,新浪网

2,1,http://news.sina.com.cn/,news.sina.com.cn,,新闻

3,2,http://news.sina.com.cn/china/,news.sina.com.cn,china,国内

```

第一列为行唯一标识号，第二列为父ID，当父ID不存在时就用-1表示。

### 2.2.2 自动新链接分类

主程序入口：eshore.cn.it.main.AutoRecognize.java

修改String url 参数值，允许本程序即可得到分类。

## 2.3 核心思想

### 2.3.1 样本关键词如何提取？

主要是通过路径分解锚链接，得到一系列关键词，比如：

http://news.sina.com.cn/china/ 那么关键词就是china

具体算法代码请看：UrlParser.java 中的 parseTextKeyWords方法。

### 2.3.2 新数据距离如何计算？

主要参考论文：基于URL特征的网页分类研究\_李玄

具体算法代码请看：LabelRecognizer.java distance 函数。

修改距离算法，可以有效提高匹配精确度和效率。

# 三、测试结果