**开题报告**

# 立论依据

互联网数据挖掘，即 Web Mining，从互联网诞生不久以后就伴随着互联网出现了。早在1996年，就已经开始着手于互联网内容识别，互联网日志分析等工具的研究。而近年来随着互联网的越来越普及，尤其是电子商务的不断发展。互联网数据挖掘越来越受到重视

当前对互联网数据挖掘的应用主要集中在电子商务方面，比较普遍而且典型的应用是通过数据挖掘找出用户消费习惯，进而根据该数据为消费者提供个性化服务，从而提升顾客满意度。但是在其他方面，譬如招聘信息挖掘，化学物质结构信息等方面也有应用。但是当前互联网信息挖掘呈现两级分化的问题：一方面是大公司将互联网数据挖掘作为一个必备工具使用，另一方面则是成为了Geek的玩具。譬如：开发人员安德鲁·沃斯（Andrew Vos）闲来无事，于是决定抓取Github（全球最大的开源代码托管网站）上抓取了100万条提交信息（commit），然后扫描分析其中哪个语言的使用者更喜欢骂脏话。但是目前互联网数据挖掘并没有渗透到普通的生活中，这与目前互联网的持续普及的趋势是格格不入的

另一方面，王磊老师等在编制山东科技版化学教材时通过对六种常见报刊媒体中出现的各化学名词分析加强中学教材与社会，生活的联系，关注社会发展的需求，学生发展的需求以及学科知识习题及其发展需要三者之间关系的动态平衡。

通过互联网信息挖掘，可以看到社会生活和科学技术的发展对于化学的强烈要求。对于教材编制者，可以更加关注和反应化学与社会技术的密切关系，贴近生活，贴近社会。对于教师而言，通过互联网信息的提取，可以更加关注到化学与生活，化学与技术之间的相关联系，更加关注化学学科的实际运用。帮助教师理解实施STS教学，使学生更容易在生活中运用化学知识解决问题。

# 研究方案

## 1．研究目标、研究内容和拟解决的关键问题

通过分析近年来新闻，科普网站中化学相关的关键词，试图找出当前与化学相关的热点，从而对教材编制，教师教学起到一定的启发性，辅助作用。尤其是对STS教学中，素材选取，以及所教授知识与社会需求，技术潮流等相互呼应等起到较好的辅助作用。

本研究的主要方法采取程序编制的方法。其中核心部分使用python实现。一些细节使用linux shell处理，某些东西还要借助Windows下特定软件实现。其中主要包括：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 实现模块 | 使用软件/套件名称 | 运行平台 | 软件/套件性质 | 相关链接 |
| 网页挖掘 | Python Scrapy | Python 爬虫 | 开源软件，使用BSD协议授权 | <https://github.com/reee/pattern> |
| 中文分词 | Python jieba | Python模块 | 开源软件，使用MIT协议授权 | <https://github.com/fxsjy/jieba> |
| 词库转换 | 深蓝词库转换 | Windows | 开源软件，使用GPL V2授权 | <https://code.google.com/p/imewlconverter/> |
| Linux Shell相关 | sort, uniq等 | Linux | 二进制文件 | —— |

为了使内容比较有代表性，以alexa排名作为重要的参考指标。其中alexa是一家专门发布网站世界排名的网站，alexa排名是目前常应用的用来评价某一网站访问量的重要指标。

从alexa上选取了中国地区排名靠前的传统互联网新闻网站，包括：

|  |  |  |  |
| --- | --- | --- | --- |
| 网站类型 | 网站名称 | 网站地址 | Alexa排名 |
| 新兴的专注新闻的门户网站 | 新浪新闻中心 | [www.sina.com.cn](http://www.sina.com.cn) | 中国区第4位 |
| 网易 | [www.163.com](http://www.163.com) | 中国区第5位 |
| 传统新闻媒体的线上站点 | 凤凰网 | [www.ifeng.com](http://www.ifeng.com) | 中国区第12位 |
| 新华网 | [www.xinhuanet.com](http://www.xinhuanet.com) | 中国区第21位 |
| 新兴科普类网站 | 果壳网 | [www.guokr.com](http://www.guokr.com) | —— |
| 科学松鼠会 | songshuhui.net | —— |

通过抓取上述网站每年与化学相关内容，进行中文分词以后，找出与化学相关的关键词，并用找出的关键词绘图。

其中需要解决的关键问题包括：

1.分析内容找出需要抓取的部分。

2.根据需求进行程序设计。

3.找出合适的关键词库用于检索。

4.关键词呈现。

## 2．拟采取的研究方法（或技术路线、实验方案）及可行性分析

网页挖掘，在Using Open Web APIs in Teaching Web Mining中被分为三类，而从本研究的目的出发。本研究属于网页内容挖掘通过获取网页中的内容，并进行分析获得有用的部分。

本研究主要路线如下：



### 网页预过滤

#### 使用关键词对网页进行预过滤

到目前为止，互联网上各种网页已经非常多了，如果直接抓取内容进行分析，那么资源消耗太大。所以本研究通过一下手段来限制网页来源：

1.只搜索限定的网站：本研究挑选了六个比较具有代表性的网站作为数据来源，其中包括：（。。。）

2.过滤网页，只下载与内容相关的网页进行分析。通过Google Custom Search API可以过滤网页，得到一定时间段，并且与题目相关的网页。

## 3．本研究的特色与创新之处

传统的教材需求分析等往往是从作者的主观意识出发。一般的STS教学材料的获取也是通过教师的主观意识判断，获取到的材料多少以及质量受限于教师的自身水平，而由于各种个人因素或者社会因素限制，通常比较单一或者过时。而目前对于STS教育的研究主要集中在理论分析部分，目前还没有发现提供任何素材来源参考的文献。

本研究通过结合计算机学科中的数据分析方法，在海量的网页中选取与生活，技术比较贴近的新闻网站和科普类网站，通过化学专业词汇关键词过滤找出与现下生活，技术相贴近的新闻，科普案例等。能够提供教师教学中的STS鲜活案例，并为化学教材改革提供参考。

## 4．预期的论文进展和成果

# 论文大纲

# 主要参考文献