# 基于互联网文本挖掘的化学热点信息研究

# 研究现状：

互联网数据挖掘自互联网诞生以后就一直是人类观察研究互联网的重要方法。始于1960s，到目前为止，对于互联网挖掘的算法已经比较成熟。当前对互联网挖掘的应用集中于下面几个方面：

1.舆情挖掘。

2.服务电商：包括但不限于：评论分析，消费习惯预测与商品推荐等。

3.某些专业化研究：互联网国学信息挖掘，互联网评论情感分析等。

本文不讨论各种挖掘方法，算法的优劣，而是使用已有成熟的开源项目构建一个完成的互联网挖掘-分析系统，目的在于提取互联网上特定网站中出现的与化学相关的相关名词，进而推断当前社会对哪些化学热点较为关注，社会对哪些化学概念的保持了持续关注，以及呈现相关化学名词之间存在的一定关联。期于让化学更贴近生活，贴近社会。服务于下列基础教育目标：

1.为STS教学素材选取提供建议

2.为教材编写，改革提供建议

3.利于学生生活化学素养的培养

# 研究方法：

本文采用的主要研究方法为程序编制，研究框架如下：

为了保持本文作为化学论文的整洁性，本文会尽量减少对程序源代码的讨论，所有本文涉及到的核心代码均已经托管到github：

<https://github.com/reee/DigWebForChemNoun>

供后续研究参考使用。

## 建立化学名词词库

从实际考虑，笔者采用了以下三个方法建立化学名词词库：

### 当前主流输入法的细胞词库

笔者选取了中国较为流行的三个输入法，从对应的输入法官方网站下载了下列与化学相关的细胞词库：

|  |  |  |
| --- | --- | --- |
| 搜狗输入法 | 百度输入法 | QQ输入法 |
| 公用小化学词库.scel | 化学化工专业词汇.bdict | 分析化学.qpyd |
| 分析化学词库.scel | 化学药品名称.bdict | 化学.qpyd |
| 化学.scel |  | 化学化工.qpyd |
| 化学专业用语.scel |  | 化学周期表.qpyd |
| 化学人名.scel |  | 化学生物学.qpyd |
| 化学助剂.scel |  | 应用化学.qpyd |
| 化学化工词汇大全.scel |  | 无机化学.qpyd |
| 化学及化工专业词汇.scel |  | 有机化学.qpyd |
| 化学品名.scel |  | 材料化学.qpyd |
| 化学小词库.scel |  | 精细化工.qpyd |
| 化学生物仪器.scel |  |  |
| 化学词汇大全.scel |  |  |
| 有机化学名词.scel |  |  |
| 煤化学.scel |  |  |
| 物理化学词汇.scel |  |  |
| 高分子专业词汇.scel |  |  |

使用深蓝词库转换（https://github.com/studyzy/imewlconverter，开源项目，免费授权）将上述细胞词库文档转化为txt文档，再将文本文档合并，使用linux shell提供的sort排序以后使用awk去重，以下的命令对ciku.txt文件进行了排序，然后去除重复项，重新输出到new\_ciku.txt：

sort ciku.txt | awk ' !x[$0]++' > new\_ciku.txt

结果为：18462条。

### 化学相关词典文档

从互联网上下载了下列词典文件：

|  |
| --- |
| 英汉汉英高分子化学词典.mdx |
| 简明化学化工专业词典.mdx |
| 英汉汉英化学及化工词典.mdx |
| 英汉化学大词典.mdx |

使用GetDict（免费软件，来源不可考，已附在userful\_bin目录）从mdx中解出txt文档，使用grep配合正则表达式提取其中的中文词条，以英汉化学大词典（导出为6-t.txt）为例，解出来的文本摘录如下：

metaarsenate <font size=5>metaarsenate\n</font>\n<br> n.偏砷酸盐\n

meta arsenate <font size=5>meta arsenate\n</font>\n<br> n.偏砷酸盐\n

metaarsenic acid <font size=5>metaarsenic acid\n</font>\n<br> n.偏砷酸\n

meta arsenic acid <font size=5>meta arsenic acid\n</font>\n<br> n.偏砷酸\n

meta arsenite <font size=5>meta arsenite\n</font>\n<br> n.偏亚砷酸盐\n

metaarsenite <font size=5>metaarsenite\n</font>\n<br> n.偏亚砷酸盐\n

meta arsenous acid <font size=5>meta arsenous acid\n</font>\n<br> n.偏亚砷酸\n

metaarsenous acid <font size=5>metaarsenous acid\n</font>\n<br> n.偏亚砷酸\n

可以很明显的发现我们需要的文本位于n.与 \n之间。通过下面的脚本可以提取出n.与 \n之间的文本，并替换n.为空，替换\n为回车并排序去重：

grep -oP 'n\..\*\\n' 6-t.txt | sed 's/n\.//g' | sed 's/\\n/\n/g' | sort | awk ' !x[$0]++'

合并并去重以后关键词数目为：17851条。

物竞化学品数据库

笔者利用scrapy爬虫从专业化学品数据库：物竞化学品数据库（http://www.basechem.org/）爬取了该数据库目前所提供的所有化学品名单，使用到的scrapy爬虫源码托管在下列地址：

[https://github.com/reee/DigWebForChemNoun/blob/master/spider/basechem/](https://github.com/reee/DigWebForChemNoun/blob/master/spider/basechem/basechem/spiders/basechem.py)

通过该化学品数据库获得的中文关键词一共7277个词条。

### 关键词清洗

将上面两个关键词源合并去重以后，人工对关键词进行了处理包括：

1. 通过阅读教材添加了大量未出现的关键词。
2. 手工添加了周期表常见元素。
3. 清除某些因为编码错误产生的字如：尸，酌等。
4. 清除之前未处理到的符号：[] ； （等。
5. 人工删除某些词语包括：

|  |  |
| --- | --- |
| **不属于化学的关键词** | 频率，图解法，构成，构架，物体 等 |
| **除化学外其他学科也很常见的词** | 实验室，密度，准确度，能量 等 |
| **某些明显错误的词** | 氧化物中，氢氧化难 等 |
| **某些语句** | 琥珀密码子即终止密码子，氯醇橡胶参见，心肝毒素如来自链球菌 等 |
| **某些过于常见的词** | 水，化学 等 |
| **某些“半”词** | 化碱，一氯 等 |

其中，下列词语虽然可能与化学不是特别相关或者可能同样属于其他学科，但因为比较具有化学色彩，经过考虑以后仍然予以保留：

富集 微克 空白对照 空白试验 水 粒子 系统误差 饱和 培养皿 炸药 羊毛

经过合并去重并人工调整代入程序的关键词数目为：23955个。

在之后的关键词提取以后，通过人工对比，发现覆盖较好。

## 抓取指定网站内容

网络爬虫（Web crawler）也叫网络蜘蛛（Web spider），蚂蚁（ant），自动检索工具（automatic indexer），或者（在FOAF软件概念中）网络疾走（WEB scutter），是一种“自动化浏览网络”的程序，或者说是一种网络机器人。

网络爬虫的基本工作原理为：

1. 从预先定义的一个URL（统一资源定位符，即我们通常所说的网页地址，网址）抓取整个页面。
2. 解析页面，如果有符合我们预先定义模板的内容，就保存到文件或写入数据库。如果有符合我们预先定义模板的URL，则继续跟踪抓取。

本文使用了scrapy（https://github.com/scrapy/scrapy）框架来爬取。

scrapy是一个基于Twisted的爬虫框架。本文在爬取网页过程中，主要使用scrapy的CrawlSpider类作整站爬取。并使用xpath选取并解出网页中的时间，和新闻正文。最终借助python代码将最终结果写入文本文档供之后的处理。爬取下列网站使用到的spider源码全部在这里列出：

<https://github.com/reee/DigWebForChemNoun/tree/master/spider>

选择的爬取网站，以及爬取的结果如下：

|  |  |  |
| --- | --- | --- |
| 网站类别 | 网站名称 | 网站地址 |
| 新兴科普类网站 | 果壳 | http://www.guokr.com/ |
| 科学松鼠会 | http://songshuhui.net/ |
| 国家级报刊 | 中国青年报电子版 | http://zqb.cyol.com/ |
| 地方报刊 | 重庆日报电子版 | http://cqrbepaper.cqnews.net/ |
| 科技类新闻站点 | Solidot | http://www.solidot.org/ |

实践中：

1. 果壳，松鼠会和solidot是典型的博客式新闻网站。文章通过时间进行归档。可以通过网站主页跟随链接轻松抓取整个站点的所有文章内容。
2. 中国青年报电子版和重庆日报电子版属于典型的报刊型站点，可以通过构造链接地址访问以前的内容（）

B2 对文章内容进行分词

本文通过调用中文结巴分词（<https://github.com/fxsjy/jieba> ）来达到比较准确的分词的目的。

结巴分词是一个成熟的python中文分词组件，包含下列算法：

* 基于Trie树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG)
* 采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合
* 对于未登录词，采用了基于汉字成词能力的HMM模型，使用了Viterbi算法

同时为了提高分词的准确性，笔者加载了之前获取的关键词列表作为jieba的用户自定义词库。在转化关键词为jieba兼容的用户自定义词库时，通过提高长词的权重来提高长词的成词率。

Jieba所规定的用户自定义词库格式为：

词汇 权重 词性

为了保证长词优先原则，转化关键词列表为jieba兼容的用户自定义词库代码如下：

<https://github.com/reee/DigWebForChemNoun/blob/master/scripts/convert_keywords_into_dict.py>

3. 从文章中提取关键词。

当前的算法基于以下猜想，即：所有含有特定汉字/汉字组合的名词都应该是化学名词（例如，所有含有钠的，我们认为都应该是属于化学名词），同时定义了一个黑名单列表，剔除某些不合理的词语（例如，我们认为含有“水”的名词应该大多数都是化学名词，但是水平，水准等显然不属于化学名词，应该剔除）和某些jieba本身因为词库/算法局限进行的错误分词（例如：含淀粉 含盐等）。

当前的提取流程如下：

1. 首先对文章进行分段/分句。为了进一步保证提取结果，采取了开源的AdvancedLangConv（https://pypi.python.org/pypi/AdvancedLangConv/0.01）将繁体中文转化为简体中文。
2. 然后调用jieba对段落/句子进行分词。这里加载之前整理的自定义词库用于提高分词的准确性。
3. 提取词语列表中的名字，若含有定义的关键词并且没有在黑名单列表中被列出来我们则认为属于合格的化学名词，写入文件。在写入文件的时候，为了方便后续作图分析，默认输出的文件为兼容与gephi的Graphviz dot文件。

其中，整理的关键词被转化为jieba自定义词库，可以通过下列地址查看：

<https://github.com/reee/DigWebForChemNoun/blob/master/keywords/user_dict.txt>

整理的关键汉字/汉字组合可以通过下列地址查看：

<https://github.com/reee/DigWebForChemNoun/blob/master/keywords/key.txt>

整理的黑名单可以通过下列地址查看：

<https://github.com/reee/DigWebForChemNoun/blob/master/keywords/blacklist.txt>

关键词提取源码可以通过下列地址查看：

<https://github.com/reee/DigWebForChemNoun/blob/master/scripts/split_and_comp.py>

该脚本主要实现了如下功能：

* 返回每个年份文章中出现的化学关键词，并且将化学关键词组织为兼容于Graphviz的dot文件格式。供作图呈现使用。

经过人工检测若干文章后发现分词效果和最终词库内含关键词均比较符合预期。

4. 关键词呈现。

本文使用gephi（https://gephi.github.io/）进行关键词呈现。

Gephi 是一款开源免费跨平台基于 JVM 的复杂网络分析软件, 其主要用于各种网络和复

杂系统。

这里需要感谢Liu Yong先生在Udemy上开设的免费课程：Gephi中文教程（[https://www.udemy.com/gephi/#/](https://www.udemy.com/gephi/" \l "/) ）让我可以快速入门。

本文通过将提取的关键词列表整理为Graphviz的dot文件格式导入gephi进行绘图。