# 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.舆情监控。通过对海量网络舆论信息进行实时的自动舆情采集，舆情分析，舆情汇总，舆情监视，识别其中的关键舆情信息，从而做到及时应急响应，为及时的正确舆论引导提供支持。

2.服务电商：通过评论分析，购物历史分析，挖掘用户消费习惯。服务于商品推荐等。

3.其他某些专业化的研究：包括但不限于互联网国学信息挖掘，互联网评论情感分析等。

忠启泉，崔允等在《新课程的理念与创新》中指出，从历史上说，课程目标应该包含三种不同的价值取向：社会本位的课程目标、学生本位的课程目标、学科本位的课程目标，化学教学除了应当服务于学生本身对知识的需求，服务于学科发展对具备相应学科知识人才的需求，还应该服务于社会，满足社会大众对基础化学常识的要求，让真正学有所成的学生能够切实的服务于社会建设，促进社会和谐发展。

本文试图利用已有的成熟的互联网文本挖掘技术构建一个较为简单的互联网化学名词文本挖掘系统。通过追踪网络中，主要是新闻类，科普类站点的化学名词，试图通过相关化学名词的自动化提取，进而推断当前社会对哪些化学名词关注度较高，社会对哪些跨学概念保持了持续关注，以及呈现一些化学名词之间的相互关联。期于让化学教育更加贴近生活实际，贴近社会需求。期于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，课程标准改革提供建议

3.在教师对学生进行生活化学素养的培养时提供一定的方向和素材。

# 研究方法：

本文采用的主要研究方法为程序编制。研究框架如图所示：

在名词词库建立的时候，使用了深蓝词库转换（https://github.com/studyzy/imewlconverter，开源项目，免费授权）进行细胞词库转换，使用GetDict（免费软件，来源不可考，已附在userful\_bin目录）从词典文件中获取文本；在爬取数据时，主要使用爬虫框架scrapy（基于python的爬虫框架，<http://scrapy.org/> 开源项目）抓取网络数据；在文本处理时，主要选取python作为主要的编程语言，使用jieba（基于python的中文分词组件 <https://github.com/fxsjy/jieba> 开源项目）进行中文分词，同时使用了某些linux shell下的文本处理工具辅助（awk sort等）文本处理；在关键词提取时，使用了AdvancedLangConv（python的pypi组件， https://pypi.python.org/pypi/AdvancedLangConv/）在图示呈现时，主要使用gephi（基于java的应用于各种网络和复杂系统的交互式可视化探索平台，<https://gephi.github.io/> 开源项目）进行图示呈现。

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、相关算法优劣问题的讨论，本文所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

需要清醒的认识到，当前的计算机自然语言处理（Natural Language Processing，简称NLP）还处在非常基础的阶段，距离让电子计算机能够真正的理解人类的语言还有很长的路要走。因此本文在进行自然语言处理时，主要通过jieba对文章进行分词处理以后，直接与既有的化学名词词库进行对比，提取出出现在化学名词词库中的名词。如何产生一个包含大多数化学名词的，贴近生活实际，且比较准确的化学名词词库，就成为本研究最重要的任务。

综合考虑，本文主要从输入法的细胞词库，网络上既有的词典文件，以及物竞化学品数据库（<http://www.basechem.org/>）三个方面建立化学名词词库。

#### 当前主流输入法的细胞词库

细胞词库是搜狗首创的、开放共享、可在线升级的细分化词库的功能名称。

细胞词库相对于的系统默认词库而言的(如下图)，其意义是满足用户的个性化输入需求。 一个细胞词库就是一个细分类别的词汇集合，细胞词库的类别可以是某个专业领域（如医学领域词库），也可以是某个地区（如北京地名词库），也可以是某个游戏（如魔兽世界词汇）。

可以发现，输入法的细胞词库为提供贴近生活的化学名词提供了一个非常优良的来源。但是必须要注意到，输入法的细胞词库是为加速用户输入而存在的，本身的词汇不一定是一个完整，专业的化学名词，因此在引入细胞词库中的化学名词以后需要人工清洗。

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 |  |  |

在获取相关文件以后，使用深蓝词库转换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工具进行，排序，然后使用awk去除重复的项目，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

经过去重以后的细胞词库提供的名词数目为：18462条。

#### 化学相关词典文档

化学相关词典显然能够提供权威专业的相关化学名词，因此从化学相关词典中获取化学名词也理应是一个非常好的途径。但是从实体词典一个个的将相关名词输入电脑效率太低。因此本文通过从互联网下载化学相关词典文件，然后解出相关的名词词汇的方式来获取相关名词。

本文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]++'

即处理后得到的文本如下：

偏砷酸盐

偏砷酸盐

偏砷酸

偏砷酸

偏亚砷酸盐

偏亚砷酸盐

合并并去除重复条目以后关键词数目为：17851条。

#### 物竞化学品数据库

物竞化学品数据库（www.basechem.org）是由上海物竞化工科技有限公司开发并维护的综合性公共化学品数据平台。截至目前，该数据库已收录基础化学品50000多种，标准品及标准物质2000多种，检测试剂盒500多种，是国内最全面的化学品数据平台。因此，从该数据库获取相关化学品信息，可以作为化学名词一个非常重要的补充。

本文利用scrapy爬虫从物竞化学品数据库抓取了该数据库目前所提供的所有化学品的中文名称，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“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 天然

某些名词与化学关联不是特别大、不主要属于化学研究范畴、或同时作为化学关键词和日常常用词，但是作为日常常用词更加普遍（例如命名、变性等）的，予以删除：

命名 计算机模拟 神经元 微生物 显微镜 光电 有效性 仪器仪表 变压器 总产量 促进作用 复合型 变性 磁场 比重 光 光线 实验 物质 基因 DNA 转基因 气味 体积 温度 颗粒 染色体 机理

通过直接提取相关网站所有名词，补充了一些词语，包括：

矿物质 化学物质 大气污染 空气污染 温室效应

温室气体 有色气体

对某些词语进行替换：

但是需要注意的是某些词，譬如聚碳酸酯缩写为PC，但PC更主要的被用来指代个人电脑，则为进行替换。

经过合并去重并人工调整代入程序的关键词数目为：23486个。

在之后的关键词提取以后，通过人工对比，发现覆盖较好。

**相似词语替换**

基于以下理由，并未作化学式向化学名称的转化：

1. 考虑到同一符号可能代替的意义相差甚大，例如He在常见网站中更被用于指代英文中的男性他，NA更多的被用作一个拟声词。另一方面，若仅对某一部分转化，其他的词语不作转化，则一定会引起出现频度的不同。

2. 在日常网站中，更多的被应用的是化学名称而不是化学符号，即使有使用符号，也一定会在文章内指明化学名称。因此不会过分影响到词频呈现。

## 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。

本文使用了基于python的网络爬虫框架scrapy（https://github.com/scrapy/scrapy）来协助爬取。

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。获得正文，时间，文章正文以后，按年/月分文件夹写入文本文档供后续分析使用。

选择的爬取网站，以及爬取的结果如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 网站类别 | 网站名称及内容 | 网站地址 | 爬取结果 |
| 新兴科普类网站 | 果壳科学人栏目，包括小组讨论等内容。 | http://www.guokr.com/ | 2010-2015年共11333篇 |
| 科学松鼠会原创栏目 | http://songshuhui.net/ | 2010-2015年共6701篇 |
| 国家级报刊 | 中国青年报电子版 | http://zqb.cyol.com/ | 2010-2015年共127558篇 |
| 地方报刊 | 重庆日报电子版 | http://cqrbepaper.cqnews.net/ | 2007-2015年共 篇 |
| 科技类新闻站点 | Solidot | http://www.solidot.org/ | 2006-2015年共33940篇 |

实际爬取策略如下：

1. 果壳，松鼠会和solidot是典型的博客式新闻网站（即文章通过时间进行归档，且新旧文章之间相互链接）。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，可以通过相应的日期构造出对应的链接地址访问到以前所有的内容。

需要注意的是solidot属于摘录类的网站，文章内容较短。

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>

对文章内容进行分词

本文采取的方案是直接使用jieba（结巴分词）进行分词，然后通过词库对比的方法提取关键词。这本身属于非常简陋，粗暴的方案。但是暂时是当前自然语言处理不够智能化的前提下的唯一的解决方案。同时，jieba能较好的完成根据词性，词义等将句子分解的任务。

jieba是一个成熟的python中文分词组件，主要基于下列算法：

* 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

Jieba支持通过加载用户自定义的词库来提高分词的准确性，因此本文加载了之前获取的关键词列表作为jieba的用户自定义词库载入jieba。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的核心代码如下：

def conv\_to\_user\_dict(keyword):

word\_weight = len(keyword) \* 10

user\_dict = keyword + " " + str(word\_weight) + " n"

return user\_dict

另外，为了防止某些词语被错误的划分，本文还建立了一个不属于化学名词的名词列表，加载到用户自定义词库，来确保某些词语正确成词，减少对正常词语的影响，该列表包括以下内容：

价值含量 低保金

信息化含量

信息含量

养老金

发展能量

基金

失业金

思想含量

技术含量

正能量

理性含量

电子书

电子产业

电子信息产业

电子学习平台

电子工程

电子文档

电子模拟

电子相框

电子胃肠镜

电子邮件

电子邮箱

电子配套产业

退还金

从文章中提取关键词

当前的提取流程如下：

1. 预先调用AdvancedLangConv将繁体中文转化为简体中文来确保关键词提取的准确性和完整性。
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. 将得到的文章对应的词语列表与既有的化学名词列表进行对比，若文章对应的词语列表在既有的化学名词列表中出现，则被认为是化学名词，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。

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>

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。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。

需要注意的是，因为提取出的关键词多而且杂，在最终呈现的时候，为了保证呈现效果，使用gephi滤掉了出小频率较低的化学名词。

局限

1. 受限于当前自然语言处理水平，在对关键词进行提取时，主要受jieba分词的准确性限制。虽然目前jieba被公认为最好的中文分词组件之一，但是仍然会出现一些误划分。导致某些关键词漏提取或者错误提取。
2. 受限于目前自然语言处理水平。暂时无法简单的区分某些词语具体语境。例如结构、命名、变性等。无法确定是否以化学名词形式出现，且作为非化学名词出现可能较多。故在操作时只能将这类词语全部排除。而比如水分、分子等，同样无法确定是否以化学名词形式出现，但作为化学名词出现应较多，故予以保留。
3. 受限于时间关系，化学名词词库完整性，准确性还有待进一步提高。
4. 虽然提取出的关键词可以较多，但是受限于人类语言本身特点，导致较为复杂关键词可能出现频率会较低，在作图时为了图示简洁，较为复杂关键词将无法得到呈现。例如今年诺贝尔化学奖得主托马斯·林达尔的研究中，出现的“水解脱嘌呤、胞嘧啶残基脱氨基、鸟嘌呤和嘧啶残基的氧化以及腺嘌呤残基甲基化生成3-甲基腺嘌呤”这些名词。因为出现频度可能较低，在作图时将会被直接忽略。
5. 某些翻译上或者语言表述上的分歧，可能导致某些关键词权重降低而不能被呈现，例如勒沙特列原理、勒夏特列原理、平衡移动原理表述的条件改变对平衡移动的影响，但是受限于当前的自然语言处理水平和笔者自身水平，时间限制，暂时无法将三整合计算。

致谢

感谢GNU，gephi，python，scrapy等开源项目维护人员的辛勤工作。

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/#/> ）让我可以快速入门使用gephi。感谢魏锐老师对我的指导。感谢学校中学教学化学组各位前辈对我的大力支持。

谢谢你们！