资料学堂

地质资料信息提取技术研究（一）

全国地质资料馆

资料服务室 郑 啸

2020年5月

引用格式请参考：郑啸.地质资料信息提取技术研究（一）(V1)：中国地质调查局发展研究中心,全国地质资料馆[创建机构],2020.全国地质资料馆[传播机构],2020-5-9. 10.37088/ngac.ppt5

地质资料包括了正文、附图、附表、附件等一系列内容，从数据角度主要可以分为两大类，一类是文本对象，包括报告文字、表格等；另一类是图件对象，包括了报告插图、各类栅格图像以及空间数据。

文本/表格

图像/空间数据

作为地质资料信息提取技术研究系列的第一讲，主要介绍**新词发现**技术。
这类对文本信息的提取和挖掘可以归入到自然语言处理的范畴。

自然语言处理(NLP)相关任务

自然语言处理应用

文本分类

文本聚类

信息检索

信息推荐

机器翻译

知识图谱

智能问答

自动摘要

情感倾向分析

多语言信息处理

语义分析

命名实体识别

关系实体抽取

指代消解

语义角色标注

.....

句法分析

依存句法分析

短语句法分析

阻块分析

.....

词法分析

分词

新词发现✓

词性标记

.....

中文的字、句、段落有明显分隔，但在词上却没有分隔符，因而在词这一层面上的识别，中文比之英文要复杂和困难得多。

分词是很多自然语言处理后续研究工作（如自动标引、知识图谱、智能检索、推荐、文本分类、自动摘要、信息挖掘等）的基础。分词结果决定后续一系列自然语言处理高级任务的分析结果，而分词的准确性本身又很大程度上依赖于词库的完整性。

因而新词发现是一个非常重要的NLP课题，也是NLP的基础任务之一。

实践中遇到的问题：

目前很多的分词工具可以较好处理大部分的通用语料，并且带有一定的未登录词（新词）发现功能，而对于网络文本、领域文档这两类文本对象的处理上，效果仍然不是很好。

这两类文本的特点在于包含大量新词，一般词典的涵盖程度比较低。而地质资料恰好又是领域文档中相对复杂的一类。

数据是有偏的

在开展地质资料相关文本研究和系统开发时的传统做法是尽可能引入相关词库，如《人民日报》语料库、搜狗词库、jieba分词词库等，稍微好一些的可能还会引入地质叙词表、地质大词典、百科词条等地质专业类词条。

但一些常规性的、新闻类、生活类词条以及其词频的介入还可能会导致一些瑕疵。如：

裤子山幅K46E015009 1/5万区域地质调查报告

裤子/n山/n幅/qK46E015009/x1/m//w5万/m区域/n地质/n调查/vn报告/n

内蒙古自治区桌子山煤田

内蒙古自治区/ns桌子/n山/n煤田/n

地质调查

地质/n调查/vn

本次介绍一种基于统计的新词发现方法，本质上属于一种典型的无监督式学习方法。即不依托于任何已有词库，在目标文本对象中，仅通过统计字与字的共现关系，进而识别出可能的词汇。

通过进一步对识别出的词汇进行清洗，进而补充原有词库，达到更好的识别效果。（数据纠偏）

裤子山幅 K46E015009 1/5万 区域地质调查 报告

↑ ↑ ↑ ↑ ↑

图幅名 图幅号 比例尺 工作类型 资料类型

资料学堂

基于统计的新词发现方法

1792档的公开版地质资料报告正文。以矿产勘查和环境灾害调查报告居多。

相关报告可以在全国馆网站—专题—非涉密地质资料专题在线浏览，或通过在线、到馆申请复制。

2.模型——马尔科夫链

x_{i+1} 的分布状态只和 x_i 有关，和其他变量条件独立，这种顺次演变的随机过程，就叫做马尔科夫链。

基于统计的自然语言处理通常都是以此为假设前提的。每一个句子是一个马尔科夫链，也就是说，句子中的每个字都只由它前一个字符决定。

3. 什么是一个词：

1. 凝固程度

全国地质资料馆

2. 自由运用程度

全国地质资料馆

(1) 凝固程度——概率统计

离散型随机变量独立性的判断 $P(A,B)=P(A) \cdot P(B)$

事件A $P(A)=0.5$
事件B $P(B)=0.25$
 $P(A, B)=0.5$

完全独立

完全不独立

$P(\text{地质}) ___ P(\text{地}) \cdot P(\text{质})$ $>>$ $>$ $\approx$ $<$ $<<$

凝固度计算结果

两字词

词	出现频率	前组词	出现频率	组合字	出现频率	新词凝固度
尴尬	2.18E-08	尴	2.18E-08	尬	2.18E-08	45924997
喧哗	2.18E-08	喧	3.27E-08	哗	2.18E-08	30616664.67
味喃	3.27E-08	味	3.27E-08	喃	3.27E-08	30616664.67
傀儡	2.18E-08	傀	3.27E-08	儡	2.18E-08	30616664.67
咆哮	4.35E-08	咆	4.35E-08	哮	4.35E-08	22962498.5
蹒跚	2.18E-08	蹒	2.18E-08	跚	4.35E-08	22962498.5
掣肘	3.27E-08	掣	3.27E-08	肘	6.53E-08	15308332.33
蚯蚓	6.53E-08	蚯	6.53E-08	蚓	6.53E-08	15308332.33
氤氲	6.53E-08	氤	6.53E-08	氲	6.53E-08	15308332.33
圉圉	6.53E-08	圉	6.53E-08	圉	6.53E-08	15308332.33
嶂岬	5.44E-08	嶂	5.44E-08	岬	7.62E-08	13121427.71
飒飒	2.18E-08	飒	4.35E-08	飒	4.35E-08	11481249.25
笈笈	2.18E-08	笈	4.35E-08	笈	4.35E-08	11481249.25
搔痒	2.18E-08	搔	2.18E-08	痒	8.71E-08	11481249.25
脍炙	9.8E-08	脍	9.8E-08	炙	9.8E-08	10205554.89
鸬鹚	8.71E-08	鸬	8.71E-08	鹚	9.8E-08	10205554.89
盱眙	8.71E-08	盱	9.8E-08	眙	8.71E-08	10205554.89
篱笆	7.62E-08	篱	8.71E-08	笆	8.71E-08	10046093.09
欺骗	2.18E-08	欺	4.35E-08	骗	5.44E-08	9184999.4
猗猗	1.09E-07	猗	1.09E-07	猗	1.09E-07	9184999.4
眶蟾	2.18E-08	眶	2.18E-08	蟾	1.2E-07	8349999.455
嗯稿	9.8E-08	嗯	1.09E-07	稿	1.09E-07	8266499.46
嘟噜	1.31E-07	嘟	1.31E-07	噜	1.31E-07	7654166.167
猖獗	1.09E-07	猖	1.42E-07	獗	1.09E-07	7065384.154
缢蛭	3.27E-08	缢	3.27E-08	蛭	1.42E-07	7065384.154
癩痢	8.71E-08	癩	1.42E-07	痢	8.71E-08	7065384.154
钥匙	6.53E-08	钥	1.09E-07	匙	8.71E-08	6888749.55
昵躺	2.18E-08	昵	3.27E-08	躺	9.8E-08	6803703.259
峥嵘	4.35E-08	峥	6.53E-08	嵘	9.8E-08	6803703.259
秸秆	4.35E-08	秸	1.09E-07	秆	6.53E-08	6123332.933

三字词

词	出现频率	前组词	出现频率	组合字	出现频率	新词凝固度
马圈圉	6.53E-08	马圈	6.53E-08	圉	6.53E-08	15306384
黑嘴鸥	6.53E-08	黑嘴	7.62E-08	鸥	2.5E-07	3422545.491
朱元璋	7.62E-08	朱元	7.62E-08	璋	3.16E-07	3166838.069
潘贵馨	2.07E-07	潘贵	2.07E-07	馨	3.81E-07	2623951.543
稍日哦	1.2E-07	稍日	1.42E-07	哦	3.27E-07	2590311.138
石嶙峋	2.83E-07	石嶙	2.94E-07	峋	4.46E-07	2156997.203
涂光炽	3.92E-07	涂光	4.14E-07	炽	4.57E-07	2071540.692
村蛤蟆	6.53E-08	村蛤	6.53E-08	蟆	5.55E-07	1800751.059
陈添铿	2.5E-07	陈添	2.83E-07	铿	5.66E-07	1562338.012
李春昱	1.96E-07	李春	4.03E-07	昱	3.38E-07	1441228.833
酸解烃	1.31E-07	酸解	1.31E-07	烃	9.15E-07	1093313.143
卤代烃	9.8E-08	卤代	9.8E-08	烃	9.15E-07	1093313.143
邱家骥	2.5E-07	邱家	9.58E-07	骥	2.5E-07	1043617.091
用玛瑙	1.31E-07	用玛	1.31E-07	瑙	1E-06	998242.4348
维尼纶	1.09E-07	维尼	1.42E-07	纶	8.82E-07	872158.6325
镇友谊	3.27E-07	镇友	3.59E-07	谊	1.06E-06	860715.1265
季凉爽	3.38E-07	季凉	3.38E-07	爽	1.22E-06	819984.8571
较偏僻	1.09E-07	较偏	2.07E-07	僻	6.42E-07	819253.3809
滴滴涕	1.03E-06	滴滴	1.22E-06	涕	1.05E-06	811443.3482
直反觚	2.07E-07	直反	2.61E-07	觚	9.8E-07	807836.9333
墨西哥	3.81E-07	墨西	3.81E-07	哥	1.26E-06	791709.5172
摩洛哥	1.52E-07	摩洛	1.52E-07	哥	1.26E-06	791709.5172
族同胞	1.31E-07	族同	1.31E-07	胞	1.3E-06	771750.4538
直返觚	1.96E-07	直返	2.61E-07	觚	9.8E-07	765319.2
聚乙烯	3.7E-07	聚乙	3.7E-07	烯	1.35E-06	740631.4839
松莫茸	2.83E-07	松莫	3.05E-07	茸	1.28E-06	722698.5182
持谨慎	9.8E-08	持谨	9.8E-08	慎	1.44E-06	695744.7273
林晓琳	1.52E-07	林晓	1.85E-07	琳	1.2E-06	687559.4952
日凌晨	6.64E-07	日凌	6.75E-07	晨	1.44E-06	684523.0381
来势凶	2.94E-07	来势	9.47E-07	凶	4.57E-07	678608.1576

(2) 自由运用程度——信息熵

香农 (Claude Shannon)最早提出了用数学方法处理自然语言的想法。

信息熵的公式: $-\log(p(x))$ ——反映事物带来的信息量

$$H(\mathbf{x}) = -\sum_{i=1}^n p(x_i) \log(p(x_i))$$

举例:

掷骰子: 1、2、3、4、5、6

信息量=信息熵 $-\log(1/6)=0.778$

掷骰子: 1、1、1、2、2、3

1信息量 $-\log(1/2)=0.3$

2信息量 $-\log(1/3)=0.47$

3信息量 $-\log(1/6)=0.78$

信息熵:

$0.3/2+0.47/3+0.78/6=0.43$

做为一个词，既要有凝固度，又要能灵活应用，能搭配很多左邻右舍，这种自用运用的程度，可以用信息熵来表示。

通过统计一个词在文章样本中出现了多少次，与其左右结合的字分别有多少种，每种组合出现了多少次，可以求出其左右信息熵：

$$H(x) = - \sum_{i=1}^n p(x_i) \log(p(x_i))$$

为表征一个词的灵活运用程度，可以选择左右信息熵中较小的值作为代表。

邻域自由度的计算结果

两字词

词	出现频率	左结合种数	右结合种数	邻域自由度
以及	0.000158	665	1105	5.424218
进入	4.45E-05	386	493	5.016289
山村	4.36E-05	372	351	5.015639
采用	0.000352	555	917	5.013007
乡的	1.25E-05	277	319	4.91947
出现	0.000184	541	672	4.890661
河和	6.95E-06	217	251	4.880356
具有	0.000203	524	719	4.879515
村下	1.45E-05	234	244	4.879019
没有	7.66E-05	370	505	4.868667
一带	0.00027	1075	624	4.848296
地区	0.000369	790	874	4.833173
也是	4.44E-05	298	486	4.812685
通过	0.000224	443	760	4.793932
全部	5.83E-05	366	412	4.783103
水库	6.83E-05	506	400	4.780005
使用	6.09E-05	381	398	4.751148
经过	4.37E-05	272	412	4.727263
村大	2.53E-05	323	283	4.692193
根据	0.000473	511	973	4.689907
镇和	1.16E-05	226	292	4.689817
还是	9.53E-06	181	244	4.667839

四字词

词	出现频率	左结合种数	右结合种数	邻域自由度
高速公路	1.96E-05	224	193	4.518815721
矿区位于	1.85E-05	155	304	4.311720431
河发源于	2.12E-06	119	99	4.276757528
的地下水	3.6E-05	249	219	4.233954669
地面沉降	4.77E-05	227	233	4.153040461
所形成的	6.6E-06	109	144	4.141428557
断裂带的	8.4E-06	123	138	4.139660237
岩溶塌陷	5.32E-05	232	219	4.069798627
矿山开采	3.94E-05	198	227	4.03829134
潜在滑坡	5.01E-05	261	156	4.009012211
矿山企业	7.79E-06	117	126	4.006951352
探矿工程	2.62E-05	155	163	3.993922307
绝大部分	1.29E-05	142	173	3.990772838
潜在崩塌	2.73E-05	155	145	3.963226301
采空塌陷	3.57E-05	200	186	3.946514717
地面塌陷	0.00021	371	372	3.942748308
在矿区内	5.72E-06	84	119	3.922909717
地下水的	7.62E-05	239	312	3.922412373
绝大多数	7.88E-06	86	136	3.920571844
生态环境	4.33E-05	177	193	3.918886133
可划分为	1.36E-05	114	169	3.913958931
航磁异常	7.19E-06	107	102	3.909053116

最终，综合考虑凝固度和自由度，加权组合得到新词确定度。在无监督（词库等背景知识信息）的情况下，发现资料文本对象中的词。当然后续还需要一些阈值筛选和脏词清洗等工作。

两字词

三字词

四字词

五字词

六字词

.....

词	新词确定度	词	新词确定度	词	新词确定度	词	新词确定度	词	新词确定度
研究	3.543404808	地裂缝	1.6654188	地质灾害	2.4834054	同位素年龄	0.6056756	省国土资源厅	1.475324365
褶皱	3.197059539	进一步	1.3210697	遥感解译	1.7271466	资源量估算	0.4396342	环境地质问题	0.840113455
影响	3.059012481	泥石流	1.2566697	地面塌陷	1.3523397	地下水污染	0.4297184	群测群防网络	0.79267052
异常	2.922358917	马圈图	0.9938517	经济损失	1.2112082	高级工程师	0.4162956	地质灾害防治	0.534032997
垃圾	2.803375623	潘贵馨	0.9877347	工作部署	1.1971839	乌拉特后旗	0.2358068	直接经济损失	0.412443071
技术	2.793498906	的影响	0.8768402	托托敦幸	1.1189902	重稀土分馏	0.2267047	有限责任公司	0.393944572
威胁	2.753925242	比例尺	0.8542771	判断矩阵	1.0788127	萨瓦亚尔顿	0.2240068	矿业有限公司	0.344381549
污染	2.692300358	涂光炽	0.8448549	萨恨托亥	1.0478281	县人民政府	0.2164235	人类工程活动	0.343194257
碰撞	2.618846033	滴滴涕	0.8385789	总工程师	0.9979018	昭平县古袍	0.2043789	轻重稀土分馏	0.286490385
进行	2.539299383	不均匀	0.7967001	地质环境	0.996068	的地质灾害	0.2018491	地质灾害隐患	0.22567393
包括	2.521467243	碳酸盐	0.7847974	都龙牙哆	0.9530774	不稳定斜坡	0.2007935	地质环境条件	0.208837501
特征	2.477207795	黑云母	0.7274221	有限公司	0.8993715	实验室承担	0.2007296	化学性质活泼	0.205115511
珊瑚	2.452420564	霍各乞	0.6915444	河西走廊	0.8345543	第四系覆盖	0.1960927	资源储量估算	0.203822876
枢纽	2.344570423	皂火壕	0.6649627	专业队伍	0.7283559	轻稀土分馏	0.1757915	地质灾害信息	0.192709575
崩塌	2.340309613	绢云母	0.606312	海百合茎	0.728286	项目任务书	0.1633982	广西英山监狱	0.175057431
信息	2.249785339	地下水	0.5723454	莫海拉亨	0.7268933	及质量评述	0.1610841	造成人员伤亡	0.173076103

目前，全国地质资料馆正在举办“第三届全国地质资料数据创新应用大赛”，其中就有自动标引方法和算法构建，并提供1200万字左右的开放文本数据集，欢迎报名参加！

- 参考资料

<http://www3.cs.stonybrook.edu/~ychoi/cse628/lecture/03-ngram.pdf>

<https://zhuanlan.zhihu.com/p/25499358>

<http://www.matrix67.com/blog/archives/5044>

资料学堂

感谢聆听!