

环境地学数据融合平台

ENVIFUSION

线上直播课

2021年8月

数据融合与数据变换

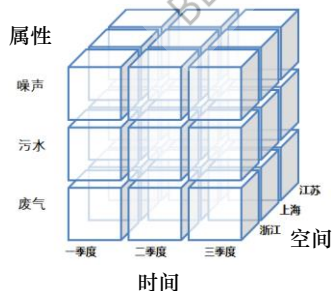
时间缺失

空间缺失

属性缺失

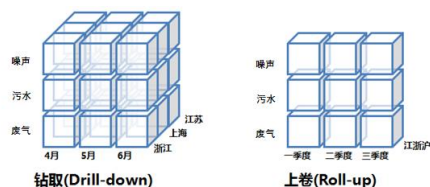
站点缺失

类型异构



原始数据

数据融合

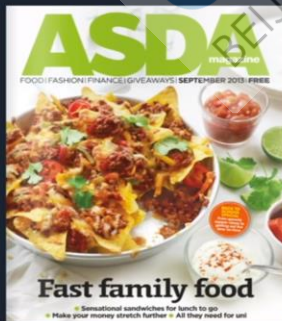


数据变换与可视化分析

异构数据

- **自然语言数据**：文字描述信息
- **图像数据**：地图、CT图像、MRI核磁图像等
- **矢量图形**：CAD、.shp场地图等
- **钻孔数据**：多为Z向一维数据

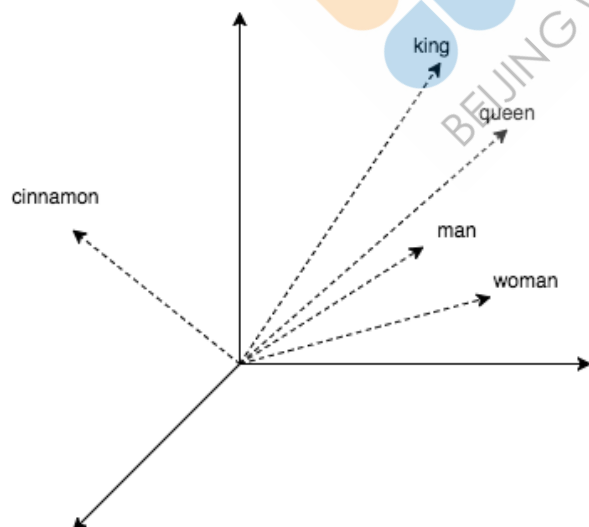
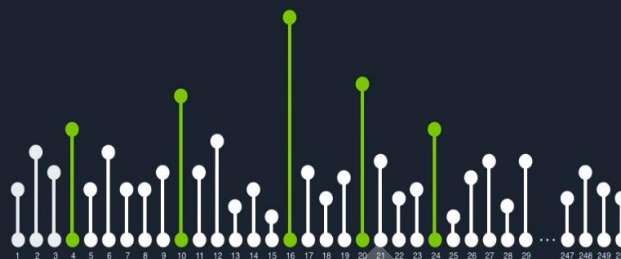
自然语言分析



```
0010100100010010101011001111
0101010100100011111010010010
1011100100101001110001010100
0010111010101010010100100010
0101010110011110101010100100
011110100100101011100100101
0011100010101000010111010101
0100101001000100101010110011
1101010101001000111110100100
1010112001001010011100010101
00001011101010100101001000
1001010101100111101010101001
000111101001001010111001001
0100111000101010000101110101
0101001010010001001010101100
111010101010010001111101001
0010101110010010100111000101
0100001011101010101000001110
0010101010111001010111010101
```



自然语言的数学表达



$$\text{Vec}(\text{'King'}) - \text{Vec}(\text{'Queen'}) + \text{Vec}(\text{'Woman'})$$

$$= \text{Vec}(\text{'Man'})$$

关键词“向量化”

基于文字的数学运算

```

命令窗口
不熟悉 MATLAB? 请参阅有关快速入门的资源。

"Tokyo"

>> word = vec2word(emb, word2vec(emb, "Paris") - word2vec(emb, "France") + word2vec(emb, "China"))
word =
    "Beijing"

>> word = vec2word(emb, word2vec(emb, "Paris") - word2vec(emb, "France") + word2vec(emb, "England"))
word =
    "London"

>> word = vec2word(emb, word2vec(emb, "Paris") - word2vec(emb, "France") + word2vec(emb, "Italy"))
word =
    "Rome"

fx >> word = vec2word(emb, word2vec(emb, "Paris") - word2vec(emb, "France") + word2vec(emb, "Japan"))

```

主题

词汇

语境



关于世界的知识



a



合适的抽象级别

环境地学专业语料大数据分析

环境地学样例语料分析

- 数据库容量与处理速度差别
- 中文与英文差别
- 语料类型差异

文本资料矢量化基础解决方案——热词矩阵

热词矩阵

自然语言由关键词（热词）组成，每一句话、每一段文字都可以看作是一个由关键词组成的向量。以环保投诉信息为例：

原始语料分词

----->1万个热词

1000
万条投诉

湖南	钢铁厂				浓烟		呛人			...
		河北		污水			呛人		恶臭	...
					...					
湖南			工地			扰民		噪声		...

热词矩阵

多个关键词向量组合起来，就会组成一个关键词矩阵，方便引入数学方法对自然语言进行处理。

创建热词矩阵

----->1万个热词

1000
万条投诉

湖南	钢铁厂	河北	工地	污水	浓烟	扰民	呛人	噪声	恶臭	...
1	1	0	0	0	1	0	1	0	0	...
0	0	1	0	1	0	0	1	0	1	...
					...					
1	0	0	1	0	0	1	0	1	0	...

热词矩阵分析路径

✓ 热词词频降维分析

- 每个词条为一行，每个热词的词频为列，将多维热词进行降维分析
- 主成分分析PCA：选取几个重要新变量进行聚类分析
- t-SNE分析：获得2-3个新变量进行空间分布展示与分类分析

✓ 热词关联挖掘

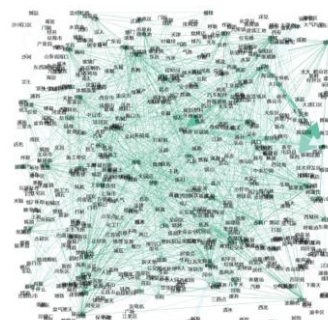
Keyword Association Mining

• KAM，热词关联挖掘：

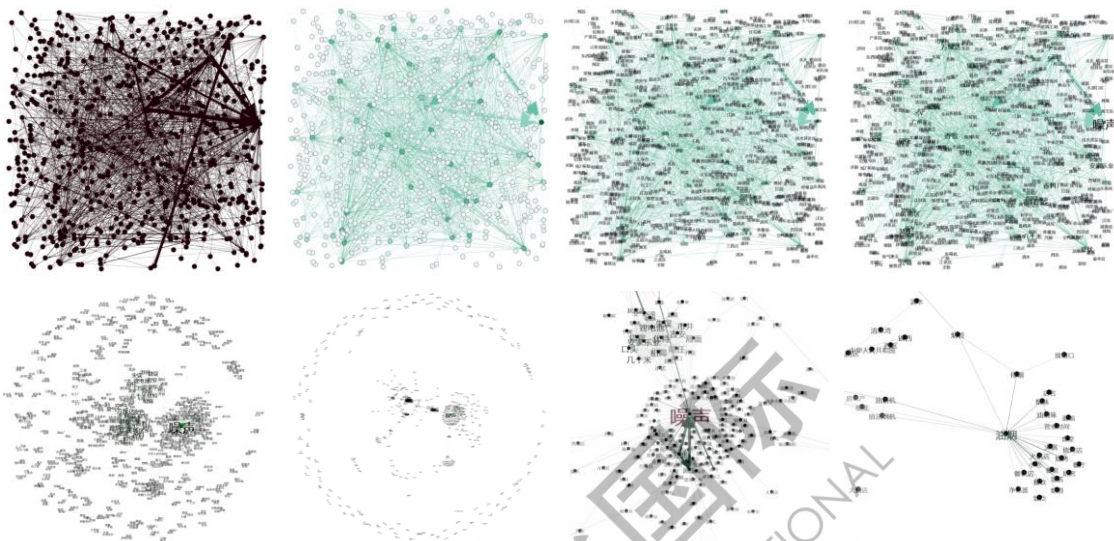
- 关联挖掘是对针对热词矩阵的一类操作，根据向量之间热词共同出现的规律来判别热词间，乃至热词向量之间的关系。由热词（节点）和热词关系（网络线）组成热词网络。

!> 700
万条关系

呛人。	浓烟。	37890。
浓烟。	恶臭。	9448。
。	...	。
钢铁厂。	呛人。	77803。



关系识别过程



专业语料分析

