1.

(1)

term	Doc 1	Doc 2	Doc 3	Doc 4
prediction	1	0	0	1
of	1	0	0	0
whole	1	0	0	0
country	1	1	0	1
sales	1	1	1	1
rise	0	1	0	1
in	0	1	1	0
July	0	1	0	1
decrease	0	0	1	0
home	0	0	1	0
June	0	0	1	0

(2)

term	doc.freq	postings list
prediction	2	1[1] 4[1]
of	1	1[1]
whole	1	1[1]
country	3	1[1] 2[1] 4[1]
sales	4	1[1] 2[1] 3[1] 4[1]
rise	2	2[1] 4[1]
in	2	2[1] 3[2]
July	2	2[1] 4[1]
decrease	1	3[1]
home	1	3[1]
June	1	4[1]

3

次序设为

```
extremly cheap DVDs CDs software thrills
```

则

```
Q= [1 2 1 1 0 0]
d1=[0 3 1 1 1 0]
d2=[0 1 1 0 0 1]
```

$$
\begin{aligned}
Q_n &= \alpha Q + \beta d_1 - \gamma d_2 \\
&= 1 \times [1\ 2\ 1\ 1\ 0\ 0] + 0.75 \times [0\ 3\ 1\ 1\ 1\ 0] - 0.25 \times [0\ 1\ 1\ 0\ 0\ 1] \\
&= [1\ 2\ 1\ 1\ 0\ 0] + [0\ 2.25\ 0.75\ 0.75\ 0.75\ 0] - [0\ 0.25\ 0.25\ 0\ 0\ 0.25] \\
&= [1\ 4\ 1.5\ 1.75\ 0.75\ -0.25]
\end{aligned}
$$

最后向量为`[1 4 1.5 1.75 0.75 0]`

4

- Manual thesaurus

人工构建同(近)义词词典

人工构建准确率高，但是费时费力

- Automatically derived thesaurus

自动导出同(近)义词词典.基于词语的共现统计信息

构建起来很迅速，但是准确性不够，冗余数据多

- Refinements based on query log mining

基于查询日志挖掘出的查询等价类

基于用户，准确率较上面要高，但是对数据要求高