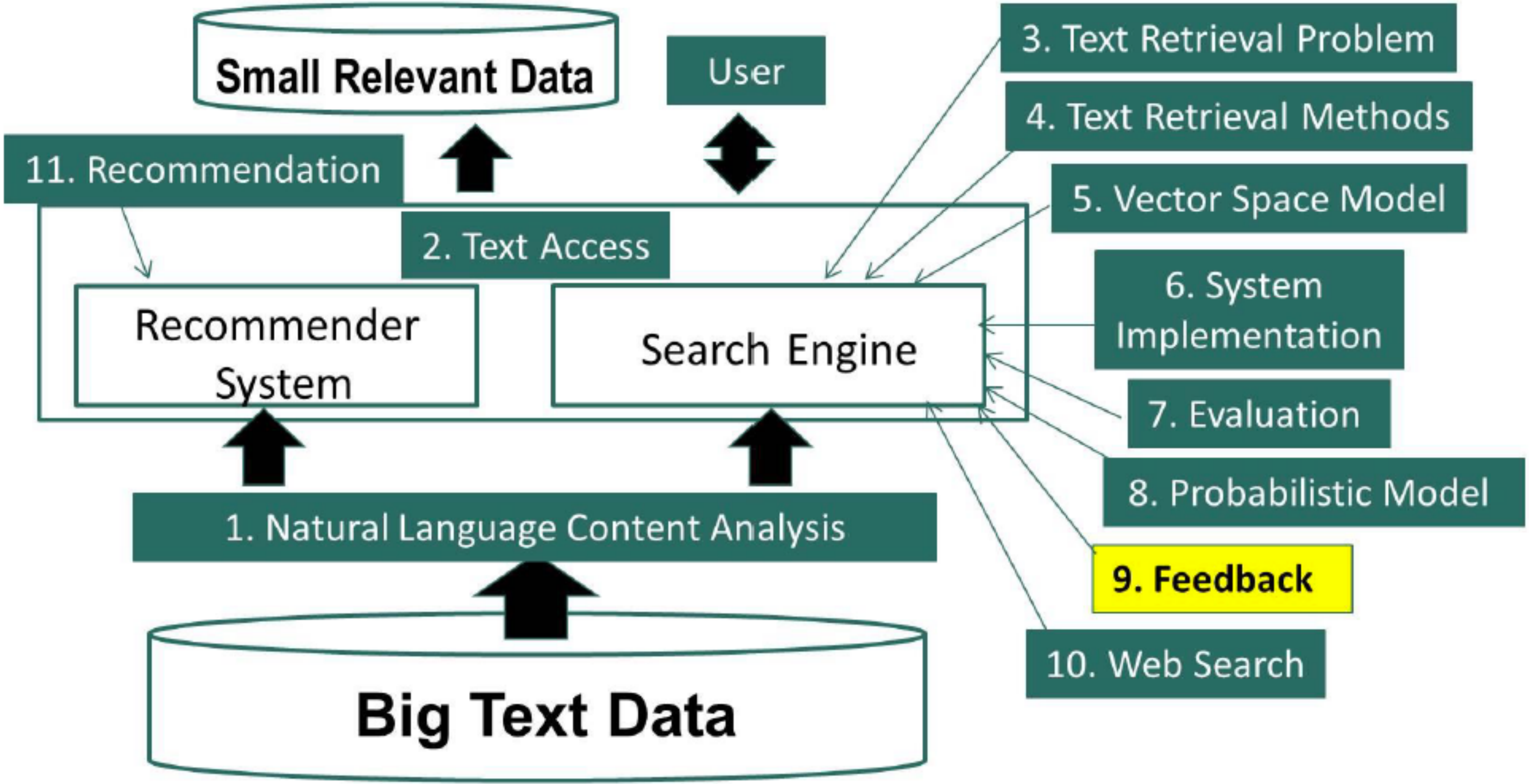
Information Retrieval & Text Mining

Retrieval Method:
Feedback in VSM

Dr. Iqra Safder
Information Technology University

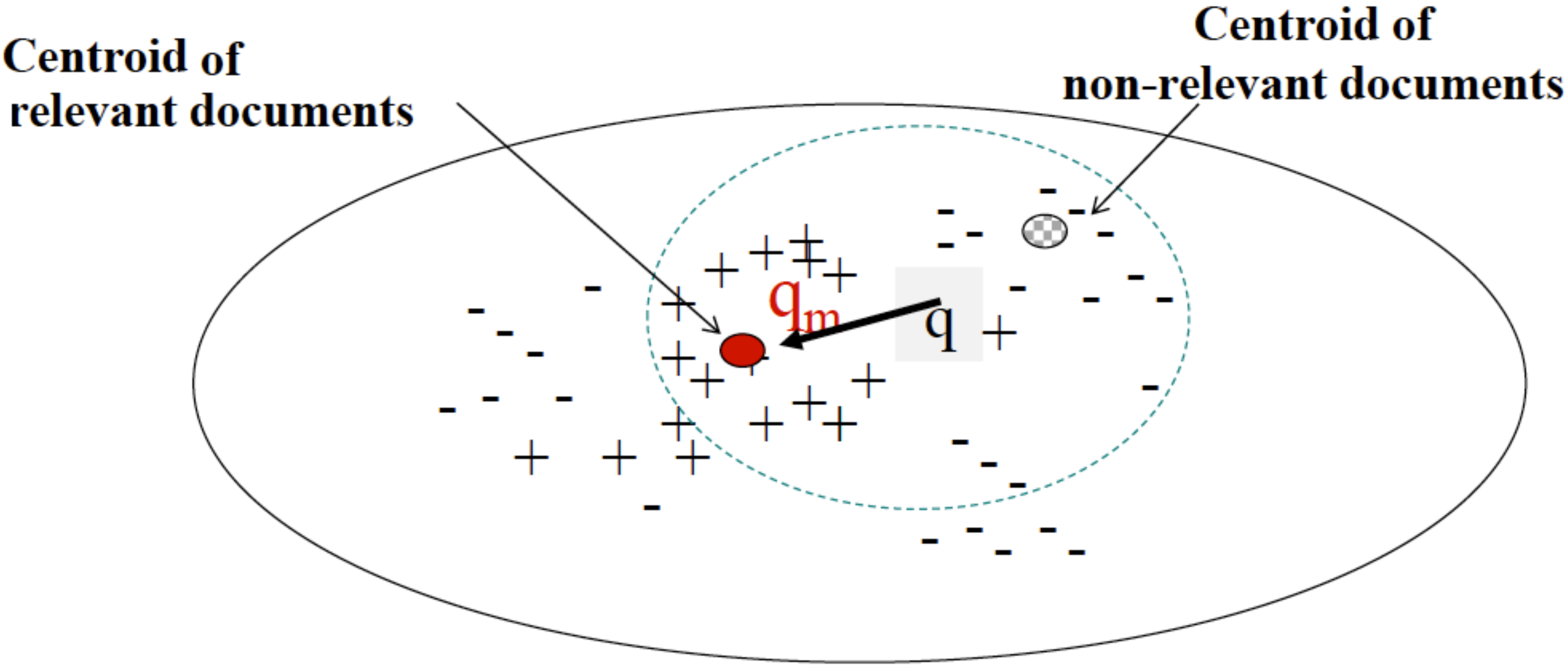
Text Retrieval Methods: Feedback in TR



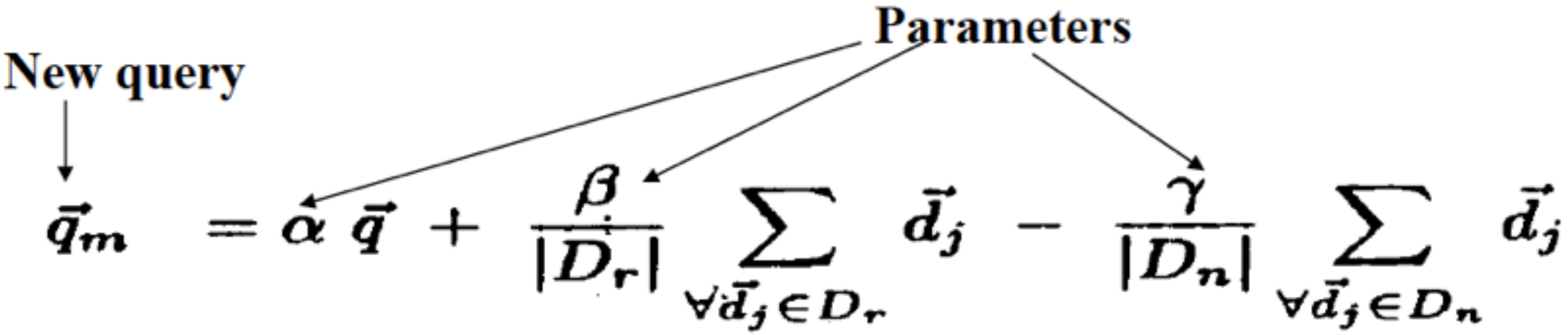
Feedback in Vector Space Model

- How can a TR system learn from examples to improve retrieval accuracy?
  - Positive examples: docs known to be relevant
  - Negative examples: docs known to be non-relevant
- General method: query modification
  - Adding new (weighted) terms (query expansion)
  - Adjusting weights of old terms

Rocchio Feedback: Illustration



Rocchio Feedback: Formula

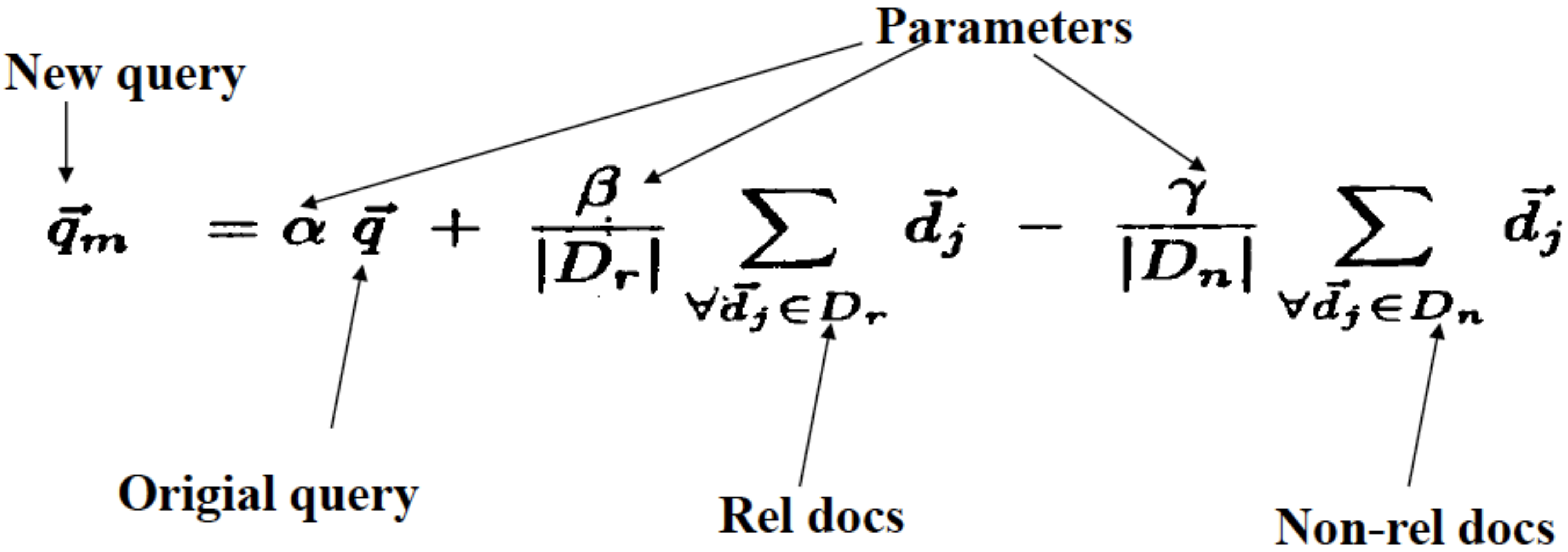


New query

Parameters

$$\vec{q}_m = \alpha \vec{q} + \frac{\beta}{|D_r|} \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|} \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

Rocchio Feedback: Formula



Parameters

New query

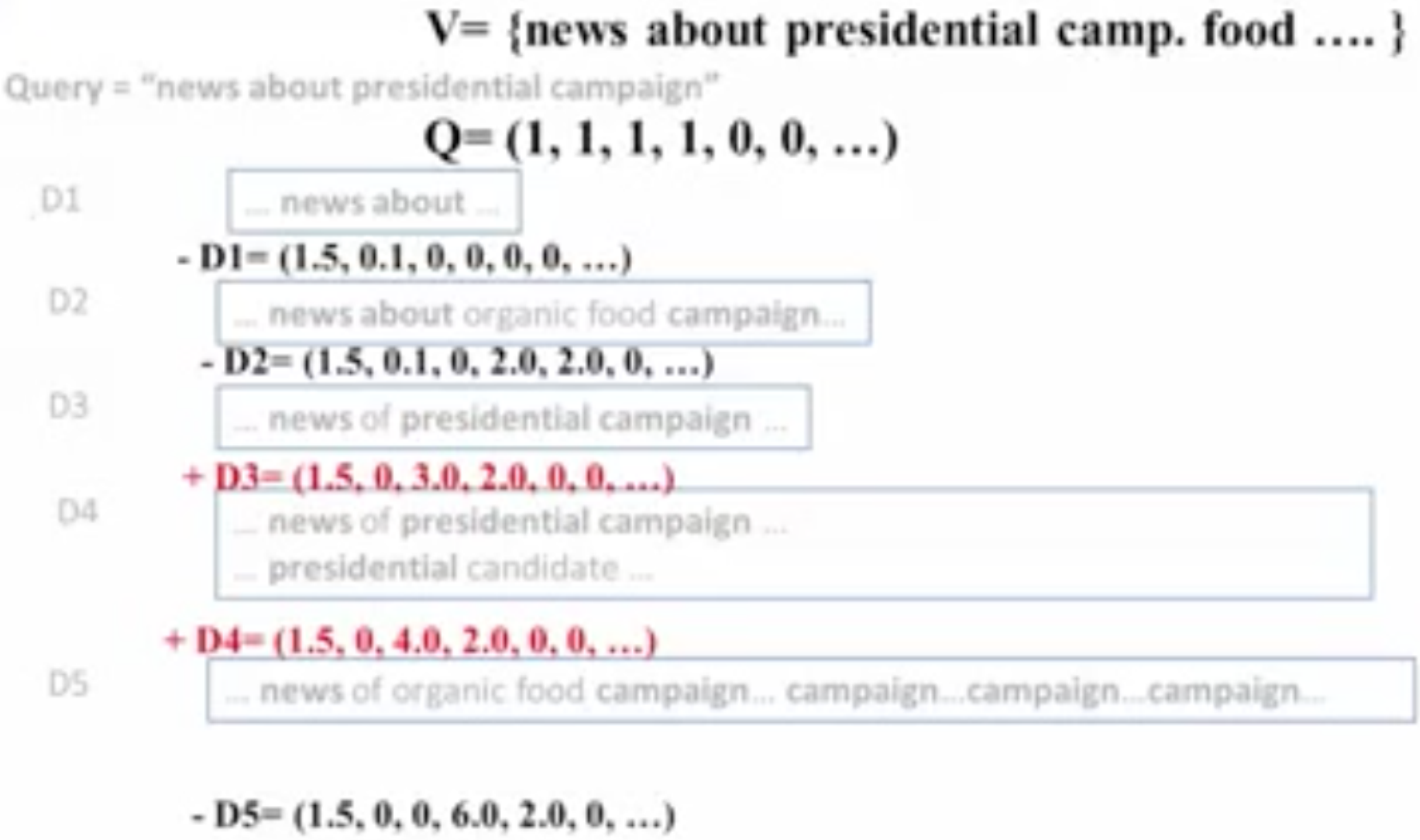
$$\vec{q}_m = \alpha \, \vec{q} + \frac{\beta}{|D_r|} \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|} \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

Origial query

Rel docs

Non-rel docs

Example of Rocchio Feedback



V= {news about presidential camp. food …. }

Query = "news about presidential campaign"

Q= (1, 1, 1, 1, 0, 0, …)

D1 … news about …

- D1= (1.5, 0.1, 0, 0, 0, 0, …)

D2 … news about organic food campaign…

- D2= (1.5, 0.1, 0, 2.0, 2.0, 0, …)

D3 … news of presidential campaign …

+ D3= (1.5, 0, 3.0, 2.0, 0, 0, …)

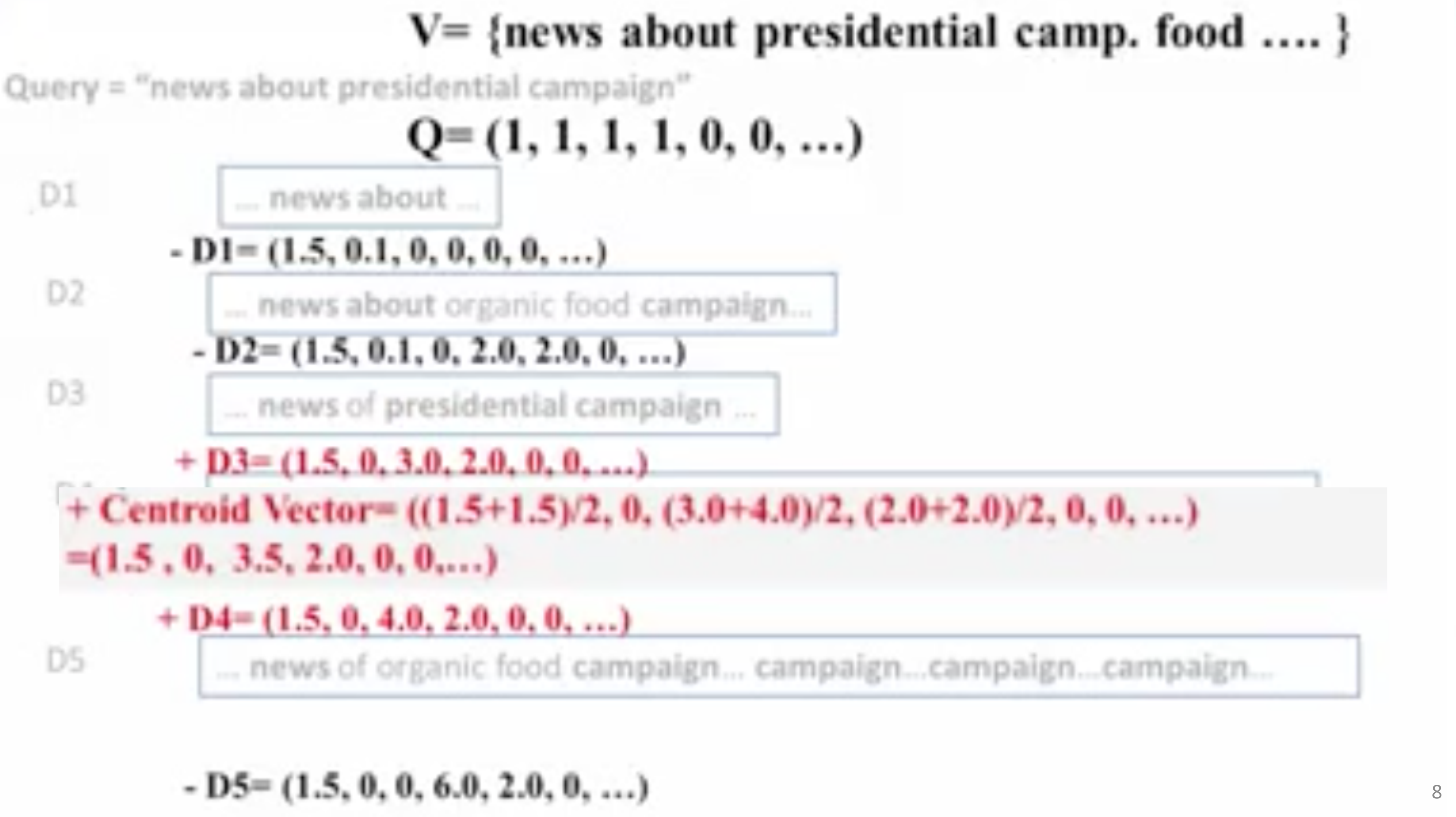
D4 … news of presidential campaign …
… presidential candidate …

+ D4= (1.5, 0, 4.0, 2.0, 0, 0, …)

D5 … news of organic food campaign… campaign…campaign…campaign…

- D5= (1.5, 0, 0, 6.0, 2.0, 0, …)

Example of Rocchio Feedback



V= {news about presidential camp. food …. }

Query = "news about presidential campaign"

Q= (1, 1, 1, 1, 0, 0, …)

D1 … news about …

- D1= (1.5, 0.1, 0, 0, 0, 0, …)

D2 … news about organic food campaign…

- D2= (1.5, 0.1, 0, 2.0, 2.0, 0, …)

D3 … news of presidential campaign …

+ D3= (1.5, 0, 3.0, 2.0, 0, 0, …)

+ Centroid Vector= ((1.5+1.5)/2, 0, (3.0+4.0)/2, (2.0+2.0)/2, 0, 0, …)
=(1.5 , 0, 3.5, 2.0, 0, 0,…)

+ D4= (1.5, 0, 4.0, 2.0, 0, 0, …)

D5 … news of organic food campaign… campaign…campaign…campaign…

- D5= (1.5, 0, 0, 6.0, 2.0, 0, …)

Example of Rocchio Feedback

V= {news about presidential camp. food …. }

Query = “news about presidential campaign”

Q= (1, 1, 1, 1, 0, 0, …)

- D1= (1.5, 0.1, 0, 0, 0, 0, …)

D2

… news about organic food campaign…

- D2= (1.5, 0.1, 0, 2.0, 2.0, 0, …)

D3

… news of presidential campaign …

+ D3= (1.5, 0, 3.0, 2.0, 0, 0, …)

D4

+ Centroid Vector= ((1.5+1.5)/2, 0, (3.0+4.0)/2, (2.0+2.0)/2, 0, 0, …)
=(1.5 , 0, 3.5, 2.0, 0, 0,…)

+ D4= (1.5, 0, 4.0, 2.0, 0, 0, …)

- Centroid Vector= ((1.5+1.5+1.5)/3, (0.1+0.1+0)/3, 0, (0+2.0+6.0)/3, (0+2.0+2.0)/3, 0, …)
=(1.5 , 0.067, 0, 2.6, 1.3, 0,…)

- D5= (1.5, 0, 0, 6.0, 2.0, 0, …)

Example of Rocchio Feedback

V= {news about presidential camp. food …. }

Query = "news about presidential campaign"

Q= (1, 1, 1, 1, 0, 0, …)

New Query Q'= (α*1+β*1.5-γ*1.5, α*1-γ*0.067, α*1+β*3.5, α*1+β*2.0-γ*2.6, -γ*1.3, 0, 0, …)

- D1= (1.5, 0.1, 0, 0, 0, 0, …)

D2

… news about organic food campaign…

- D2= (1.5, 0.1, 0, 2.0, 2.0, 0, …)

D3

… news of presidential campaign …

+ D3= (1.5, 0, 3.0, 2.0, 0, 0, …)

D4

+ Centroid Vector= ((1.5+1.5)/2, 0, (3.0+4.0)/2, (2.0+2.0)/2, 0, 0, …)
=(1.5 , 0, 3.5, 2.0, 0, 0,…)

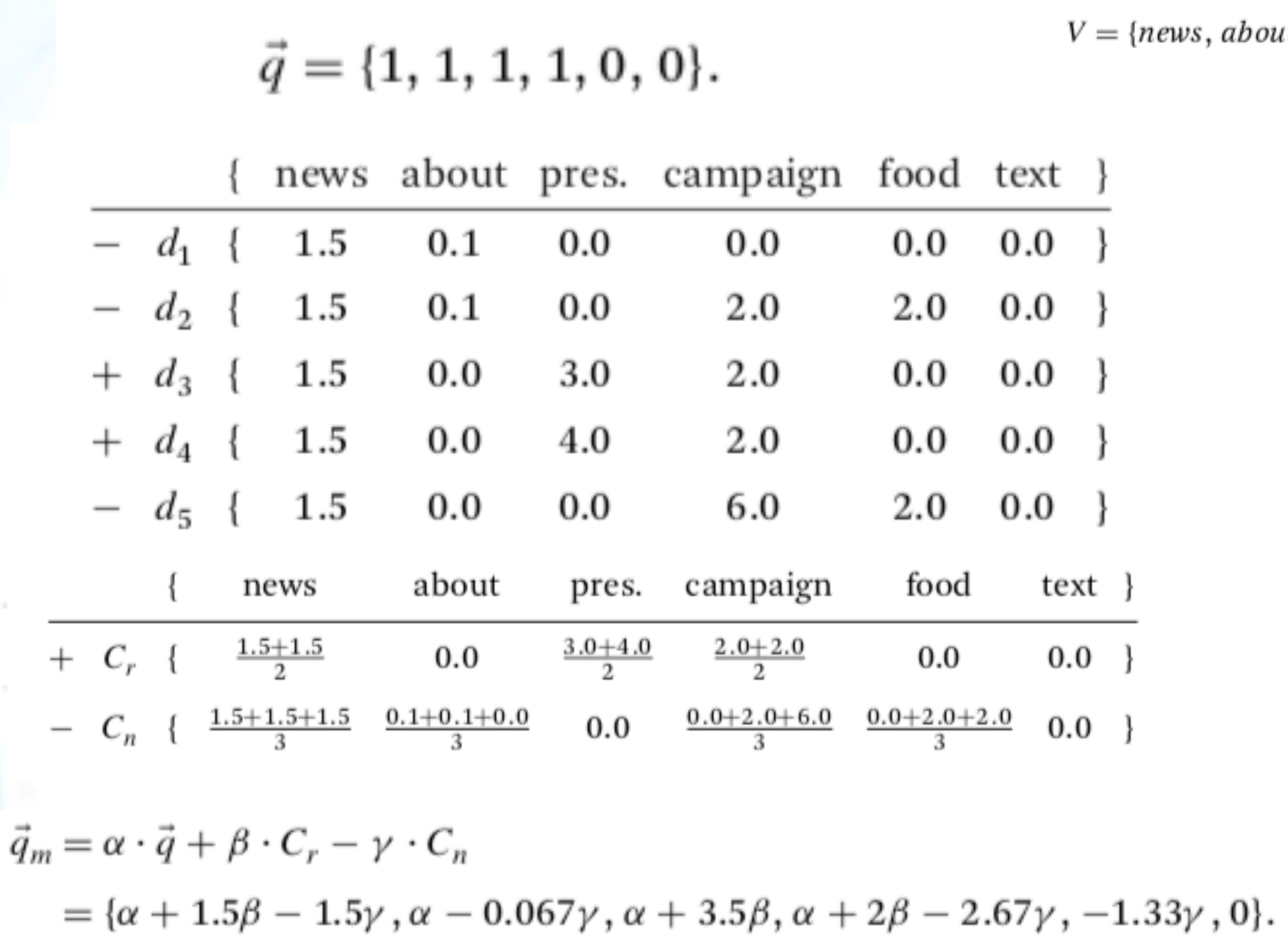
+ D4= (1.5, 0, 4.0, 2.0, 0, 0, …)

- Centroid Vector= ((1.5+1.5+1.5)/3, (0.1+0.1+0)/3, 0, (0+2.0+6.0)/3, (0+2.0+2.0)/3, 0, …)
=(1.5 , 0.067, 0, 2.6, 1.3, 0,…)

- D5= (1.5, 0, 0, 6.0, 2.0, 0, …)

After query expansion there are many non zero terms in the query vector, while the original had only 4 non zero terms. Practically we truncate long vectors and only use the terms that have higher weights.

$V = \{news, about, presidential, campaign, food, text\}$



$$\vec{q} = \{1, 1, 1, 1, 0, 0\}.$$

		{	news	about	pres.	campaign	food	text	}
−	d_1	{	1.5	0.1	0.0	0.0	0.0	0.0	}
−	d_2	{	1.5	0.1	0.0	2.0	2.0	0.0	}
+	d_3	{	1.5	0.0	3.0	2.0	0.0	0.0	}
+	d_4	{	1.5	0.0	4.0	2.0	0.0	0.0	}
−	d_5	{	1.5	0.0	0.0	6.0	2.0	0.0	}

		{	news	about	pres.	campaign	food	text	}
+	C_r	{	$\frac{1.5+1.5}{2}$	0.0	$\frac{3.0+4.0}{2}$	$\frac{2.0+2.0}{2}$	0.0	0.0	}
−	C_n	{	$\frac{1.5+1.5+1.5}{3}$	$\frac{0.1+0.1+0.0}{3}$	0.0	$\frac{0.0+2.0+6.0}{3}$	$\frac{0.0+2.0+2.0}{3}$	0.0	}

$$\begin{aligned}\vec{q}_m &= \alpha \cdot \vec{q} + \beta \cdot C_r - \gamma \cdot C_n \\ &= \{\alpha + 1.5\beta - 1.5\gamma, \alpha - 0.067\gamma, \alpha + 3.5\beta, \alpha + 2\beta - 2.67\gamma, -1.33\gamma, 0\}.\end{aligned}$$

Rocchio in Practice

- Negative (non-relevant) examples are not very important (why?) Distract the query
- Often truncate the vector (i.e., consider only a small number of words that have highest weights in the centroid vector) (efficiency concern)
- Avoid "over-fitting" (keep relatively high weight on the original query weights) (why?)
- Can be used for relevance feedback and pseudo feedback (β should be set to a larger value for relevance feedback than for pseudo feedback)
- Usually robust and effective

It's also important to avoid over-fitting, which means we have to keep relatively high weight α on the original query terms. We don't want to overly trust a small sample of documents and completely reformulate the query without regard to its original meaning. Those original terms are typed in by the user because the user decided that those terms were important! Thus, we bias the modified vector towards the original query direction.

Term Project

- 3-4 group members
- Use any textual datasets.
  - Keggal
  - ACL
  - Github
  - Google Dataset
  - Elseveir (dataset along with papers are available)
  - Paperswithcode

- Deadline to choose project ideas 18th November, 2021
- Please fill the online excel sheet.