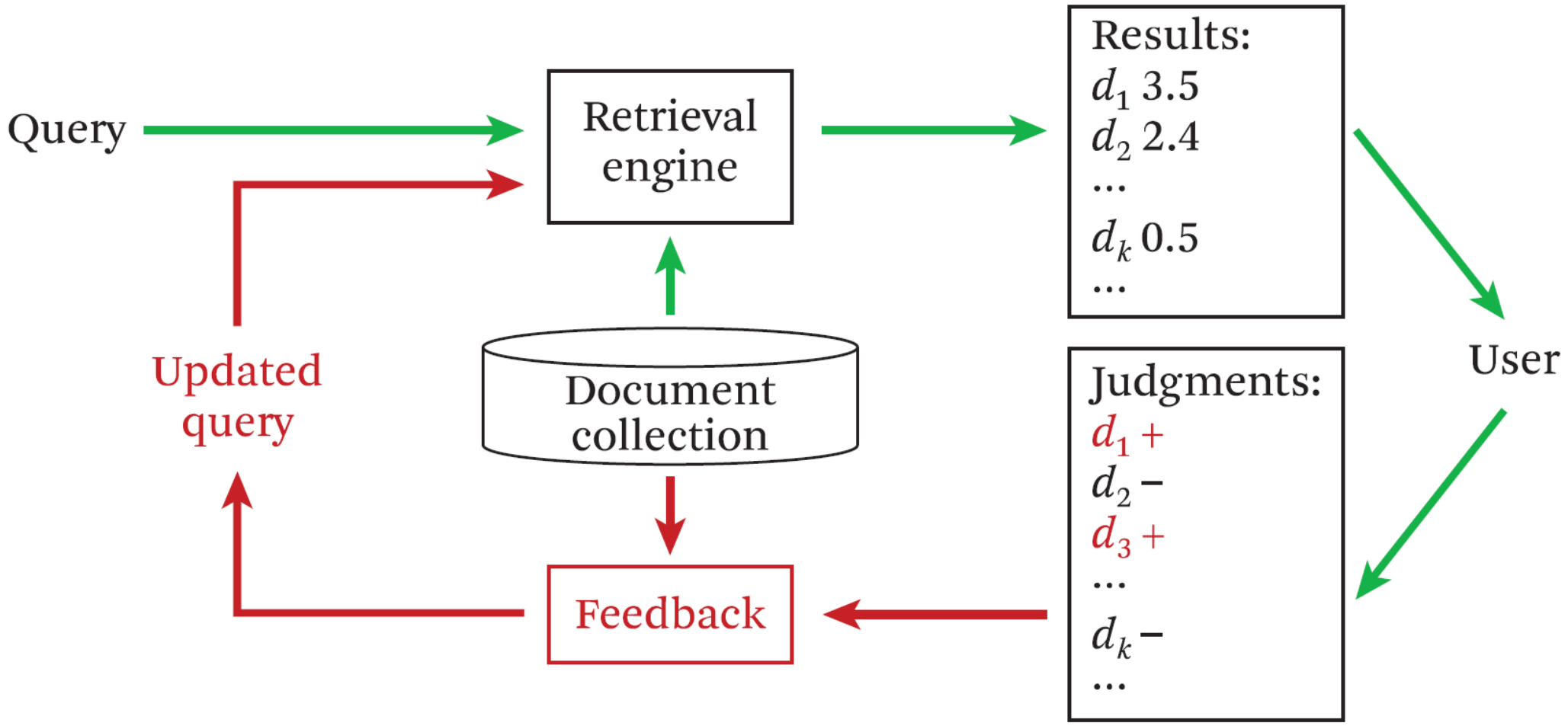
Feedback

지능정보시스템연구실

컴퓨터과학과 천상진(csjin75@uos.ac.kr)



Feedback takes the results of a user's actions or previous search results to improve retrieval results

Intro

- ✓ ***relevance feedback***

  use explicit relevance judgement by user effort

- ✓ ***pseudo feedback***

  simply assume the top k documents are relevant

- ✓ ***implicit feedback***

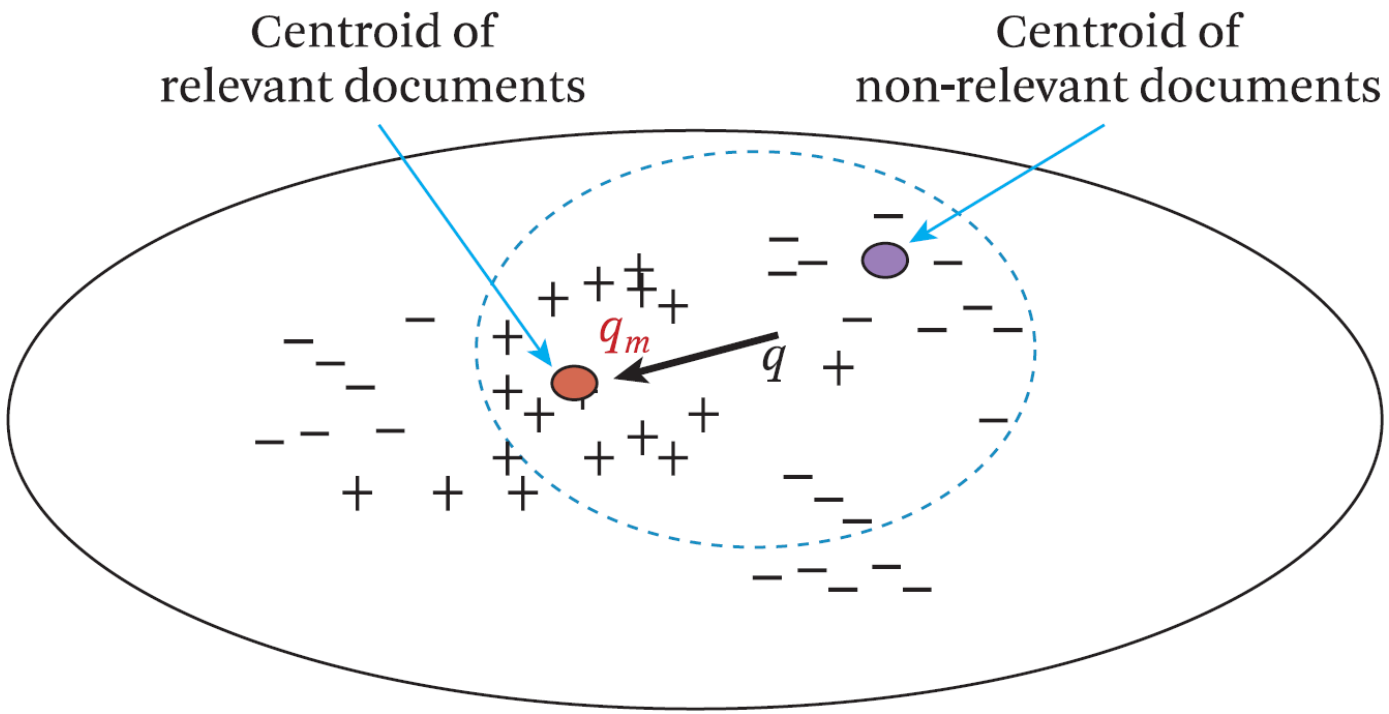
  use user's clickthrough data

*"**feedback** in a TR system is based on learning from previous queries to improve retrieval accuracy in the future queries."*

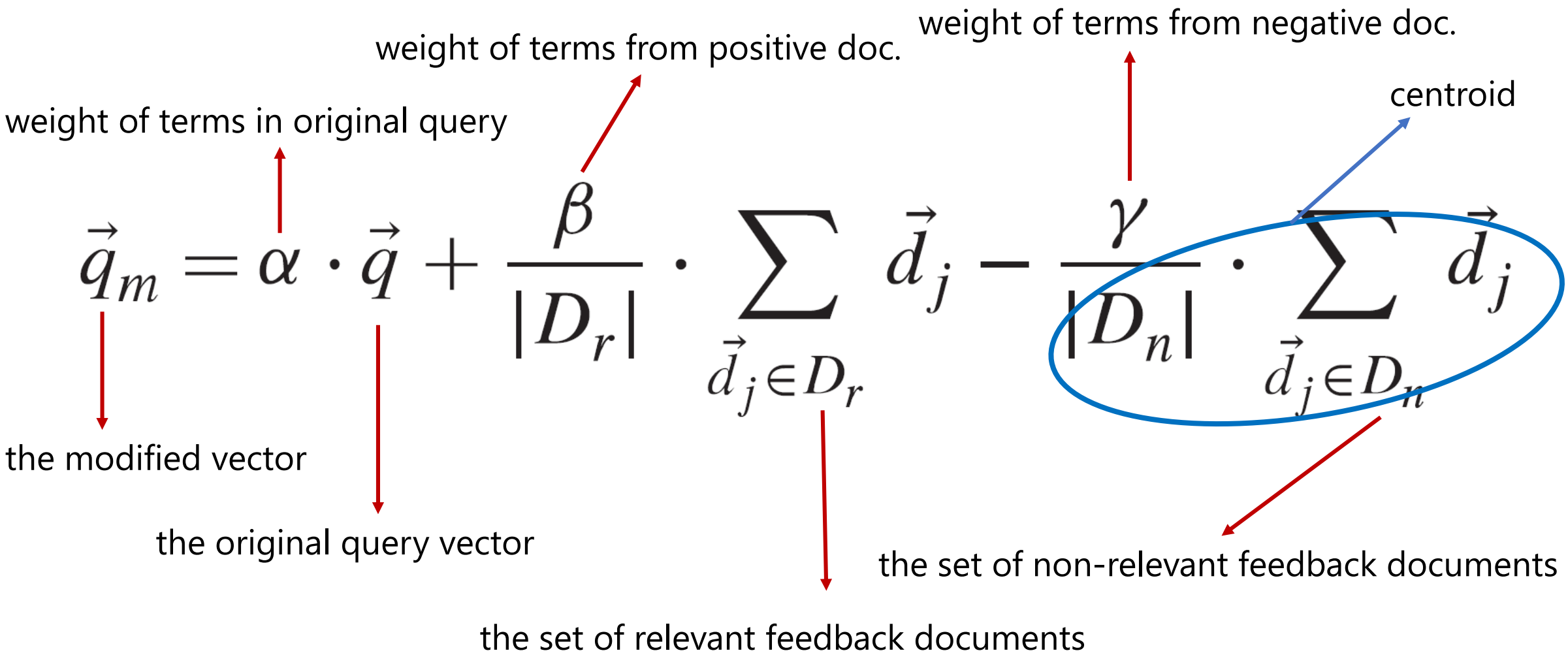
7.1 Feedback in the Vector Space Model

we want to place the query vector in a better position
in the high-dimensional term space, plotting it closer to relevant documents.

→ **Rocchio Feedback Method**



7.1 Feedback in the Vector Space Model



7.1 Example

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

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

		{	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	}

7.1 Example

✓ ***compute the centroid of positive and negative feedback documents***

		{	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	}

✓ ***modify the original query to create the expanded query*** $\vec{q}_m$

$$\vec{q}_m = \alpha \cdot \vec{q} + \beta \cdot C_r - \gamma \cdot C_n$$

$$= \{\alpha + 1.5\beta - 1.5\gamma, \alpha - 0.067\gamma, \alpha + 3.5\beta, \alpha + 2\beta - 2.67\gamma, -1.33\gamma, 0\}$$

7.2 Feedback in Language Model

Kullback-Leibler divergence retrieval model

Query likelihood

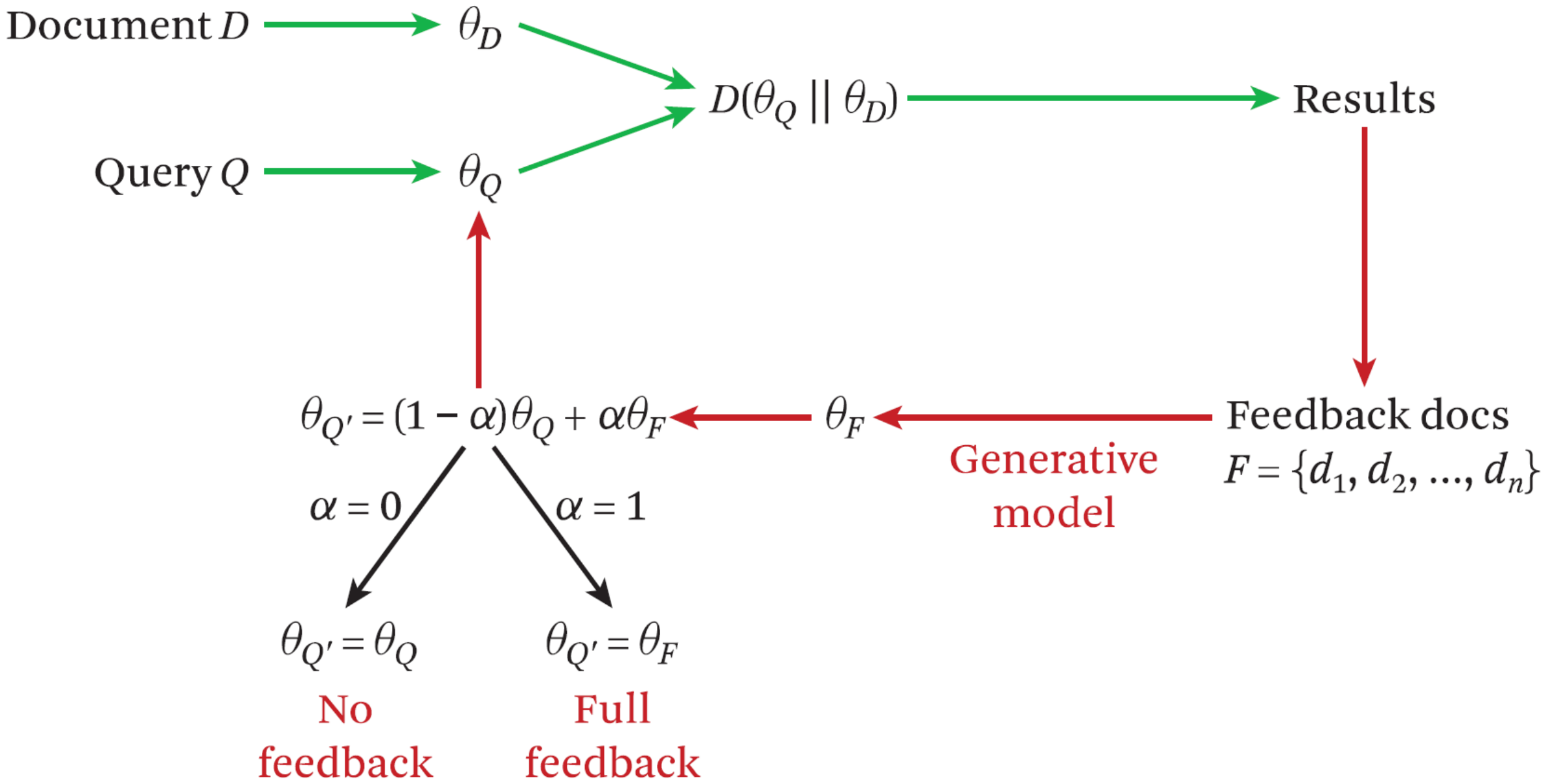
$$f(q,d)=\sum_{\substack{w\in d\\ w\in q}} c(w,q)\left[\log\frac{p_{\text{seen}}(w|d)}{\alpha_d p(w|C)}\right]+n\log\alpha_d$$

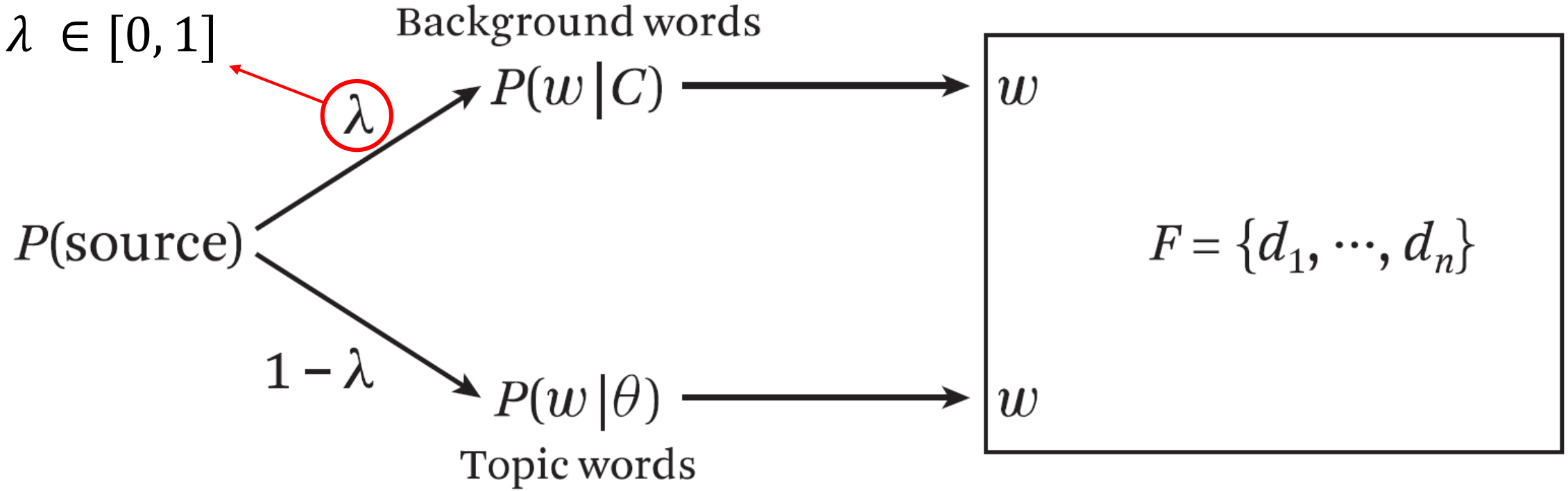
KL-divergence (cross entropy)

$$f(q,d)=\sum_{w\in d,p(w|\theta_Q)>0}\left[p(w|\hat{\theta}_Q)\log\frac{p_{\text{seen}}(w|d)}{\alpha_d p(w|C)}\right]+\log\alpha_d$$

Query LM

$$p(w|\hat{\theta}_Q)=\frac{c(w,Q)}{|Q|}$$


Document D
θ_D
Query Q
θ_Q
$D(\theta_Q \,||\, \theta_D)$
Results
Feedback docs
$F = \{d_1, d_2, ..., d_n\}$
Generative model
θ_F
$\theta_{Q'} = (1 - \alpha)\theta_Q + \alpha\theta_F$
$\alpha = 0$
$\alpha = 1$
$\theta_{Q'} = \theta_Q$
$\theta_{Q'} = \theta_F$
No feedback
Full feedback



$$\log p(F|\theta) = \sum_i \sum_w c(w; d_i) \log[(1-\lambda)p(w|\theta) + \lambda p(w|C)]$$

Maximum likelihood $\theta_F = \underset{\theta}{\text{argmax}} \log p(F|\theta)$

Query: "airport security"

Mixture model approach

$\lambda = 0.9$

w	$P(w \mid \theta_F)$
security	0.0558
airport	0.0546
beverage	0.0488
alcohol	0.0474
bomb	0.0236
terrorist	0.0217
author	0.0206
license	0.0188
bond	0.0186
counter-terror	0.0173
terror	0.0142
newsnet	0.0129
attack	0.0124
operation	0.0121
headline	0.0121

Web database

Top 10 docs

$\lambda = 0.7$

w	$P(w \mid \theta_F)$
the	0.0405
security	0.0377
airport	0.0342
beverage	0.0305
alcohol	0.0304
to	0.0268
of	0.0241
and	0.0214
author	0.0156
bomb	0.0150
terrorist	0.0137
in	0.0135
license	0.0127
state	0.0127
by	0.0125

7.2 Feedback in Language Model

$$\theta_F = \arg\max_{\theta} \log\, p(F \mid \theta)$$

$$= \arg\max_{\theta} \sum_{d \in F} \sum_{w} c(w, d) \cdot \log\left[(1-\lambda) \cdot p(w \mid \theta) + \lambda \cdot p(w \mid C)\right]$$

→ *choose θ_F to maximize the log likelihood of the feedback documents*

In practice, it isn't feasible to try all values of θ, so we use EM algorithm for estimating its parameters (Chapter 17…)