

Document Priors

Jaime Arguello

INLS 509: Information Retrieval

jarguell@email.unc.edu

Outline

Introduction to language modeling

Language modeling for information retrieval

Query-likelihood Retrieval Model

Smoothing

Priors

Linear Interpolation

Review

$$\text{score}(Q, D) = \prod_{i=1}^n (\lambda P(q_i|D) + (1 - \lambda)P(q_i|C))$$

- $P(q_i|D)$ = probability given to query term q_i by the document language model
- $P(q_i|C)$ = probability given to query term q_i by the collection language model

Linearly Interpolated Smoothing

Review

- Doc 1: haikus are easy
- Doc 2: but sometimes they don't make sense
- Doc 3: refrigerator
- Query: haikus make sense

$$score(Q, D) = \prod_{i=1}^n (\lambda P(q_i|D) + (1 - \lambda)P(q_i|C))$$

(source: threadless t-shirt)

Let's Take A Step Back

- The query likelihood model has a more theoretic motivation than I've portrayed so far



Bayes' Law

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$



Bayes' Law

A photograph of a blue-lit display board showing the formula for Bayes' Law. The formula is written in a stylized, handwritten font. The left side of the equation is $P(A|B)$, followed by an equals sign. The right side is a fraction with $P(B|A)P(A)$ in the numerator and $P(B)$ in the denominator. A horizontal line separates the numerator and denominator. The display board is mounted on a wall, and the background is dark.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

(source: wikipedia)



Bayes' Law Derivation

$$P(A, B) = P(A|B) \times P(B)$$

$$P(A, B) = P(B|A) \times P(A)$$

$$P(A, B) = P(A, B)$$

$$P(A|B) \times P(B) = P(B|A) \times P(A)$$

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$



Bayes' Law Applied to Ranking

$$P(D|Q) = \frac{P(Q|D) \times P(D)}{P(Q)}$$



Bayes' Law Applied to Ranking

$$P(D|Q) = \frac{P(Q|D) \times P(D)}{P(Q)}$$

If we're scoring and ranking documents based on this formula, which number doesn't matter?

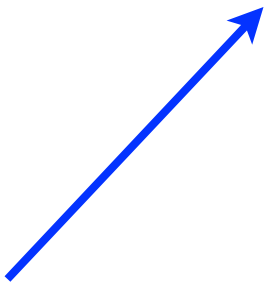
Query-likelihood Retrieval Model

- Dividing every document score by the same number doesn't change the ranking of documents ...
- So, we can ignore the denominator $P(Q)$

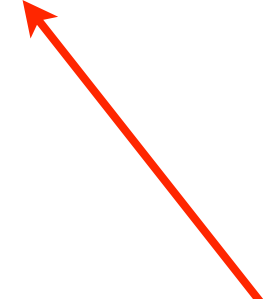
$$P(D|Q) = \frac{P(Q|D) \times P(D)}{P(Q)}$$

$$P(D|Q) \propto P(Q|D) \times P(D)$$

query-likelihood score
(you already know this)



document prior
(new concept)



Document Prior

$$P(D|Q) \propto P(Q|D) \times P(D)$$

query-likelihood score
(you already know this)



document prior
(new concept)



- The document prior, $P(D)$, is the probability that the document is relevant to any query
- It is a document-specific probability
- It is a query-independent probability

Document Prior

$$P(D|Q) \propto P(Q|D) \times P(D)$$

query-likelihood score
(you already know this)

document prior
(this is a new concept)

- Unknowingly, so far we've assumed that $P(D)$ is the same for all documents
- Under this assumption, the ranking is based only on the query-likelihood given the document language model
- Now, we will assume that $P(D)$ is not uniform
- That is, some documents are more likely to be relevant independent of the query

Document Prior

$$P(D|Q) \propto P(Q|D) \times P(D)$$

- What is it?
- Anything that affects the likelihood that a document is relevant to any query
 - ▶ document popularity
 - ▶ document authority
 - ▶ amount of content (e.g., length)
 - ▶ topical cohesion
 - ▶ really, you decide ...

Document Prior

$$P(D|Q) \propto P(Q|D) \times P(D)$$

- But, it is a probability, so in a collection of M documents...

$$\sum_{i=1}^M P(D_i) = ?$$



Document Prior

$$P(D|Q) \propto P(Q|D) \times P(D)$$

- Not that difficult...

$$P(D_j) = \frac{\textit{score}(D_j)}{\sum_{i=1}^M \textit{score}(D_i)}$$

Document Prior

$$P(D|Q) \propto P(Q|D) \times P(D)$$

- What is it?
- Anything that affects the likelihood that a document is relevant to any query
 - ▶ document popularity
 - ▶ document authority
 - ▶ amount of content (e.g., length)
 - ▶ topical cohesion
 - ▶ really, you decide ...

Document Popularity

- Given user-interaction data, we can determine the popularity of a document based on clicks
- Click-rate:

$$\frac{\text{\# of clicks on the document}}{\text{\# of clicks on any document}}$$

Document Popularity

most clicked urls - aol query-log (2006)

rank	URL	P(URL)	rank	URL	P(URL)
1	http://www.google.com	0.0204	11	http://www.geocities.com	0.0022
2	http://www.myspace.com	0.0093	12	http://www.hotmail.com	0.0022
3	http://mail.yahoo.com	0.0090	13	http://www.ask.com	0.0021
4	http://en.wikipedia.org	0.0066	14	http://www.bizrate.com	0.0017
5	http://www.amazon.com	0.0056	15	http://www.tripadvisor.com	0.0017
6	http://www.mapquest.com	0.0054	16	http://www.msn.com	0.0017
7	http://www.imdb.com	0.0053	17	http://profile.myspace.com	0.0016
8	http://www.ebay.com	0.0044	18	http://www.craigslist.org	0.0015
9	http://www.yahoo.com	0.0030	19	http://disney.go.com	0.0015
10	http://www.bankofamerica.com	0.0027	20	http://cgi.ebay.com	0.0015

Document Popularity

least clicked urls - aol query-log (2006)

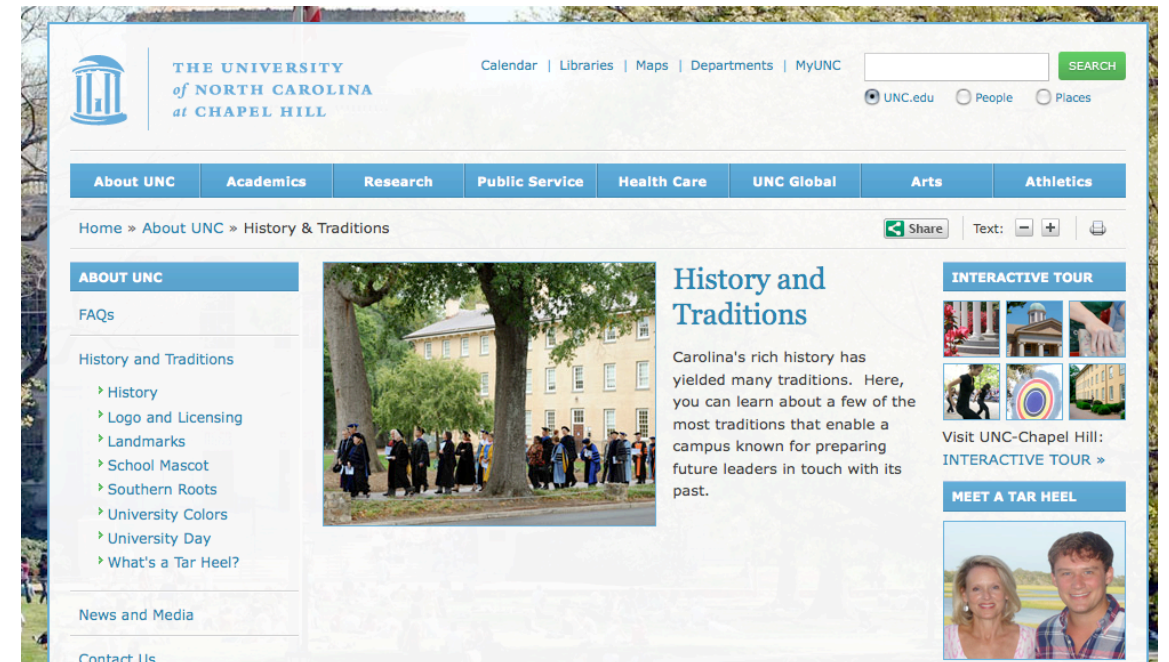
rank	URL	P(URL)	rank	URL	P(URL)
1501087	http://www.live4soccer.com	0.0000	1501097	http://www.toymod.com	0.0000
1501088	http://www.smalltowngallery.com	0.0000	1501098	http://www.aaabarcodes.com	0.0000
1501089	http://1239.8wmc5l.info	0.0000	1501099	http://www.stubaidirect.com	0.0000
1501090	http://silverjews.lyrics-online.net	0.0000	1501100	http://rtbknox.no-ip.biz	0.0000
1501091	http://www2.glenbrook.k12.il.us	0.0000	1501101	http://www.panontheweb.com	0.0000
1501092	http://www.palmerschools.org	0.0000	1501102	http://4395.bsxf57.info	0.0000
1501093	http:// www.rainbowridgefarmequestriancenter.com	0.0000	1501103	http://www.calco.com	0.0000
1501094	http://mncable.net	0.0000	1501104	http://www.sharpe.freshair.org	0.0000
1501095	http://www.modem-software.com	0.0000	1501105	http://www.opium.co.za	0.0000
1501096	http://www.clevelandrugby.com	0.0000	1501106	http://grediagnostic.ets.org	0.0000

Document Popularity

<http://www.unc.edu>



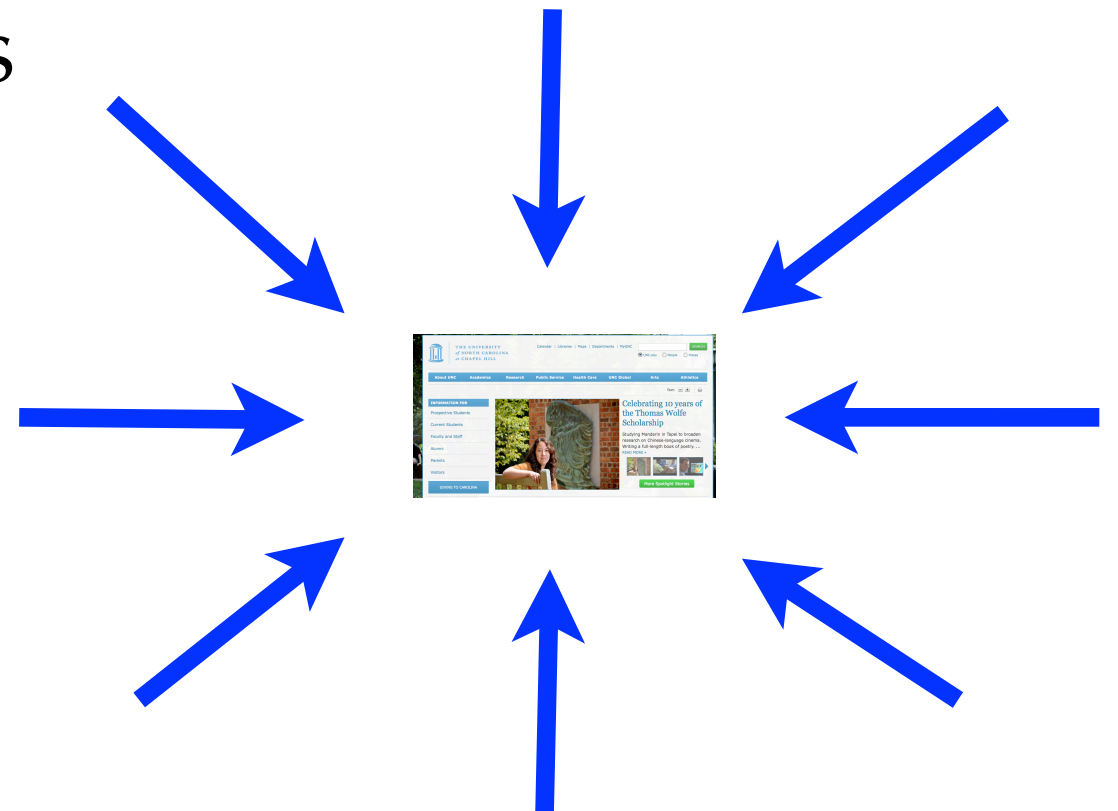
<http://www.unc.edu/about/history-traditions>



- URL depth
 - ▶ website entry-pages tend to be more popular than those that are deep within the domain
- Count the number of “/” in the URL

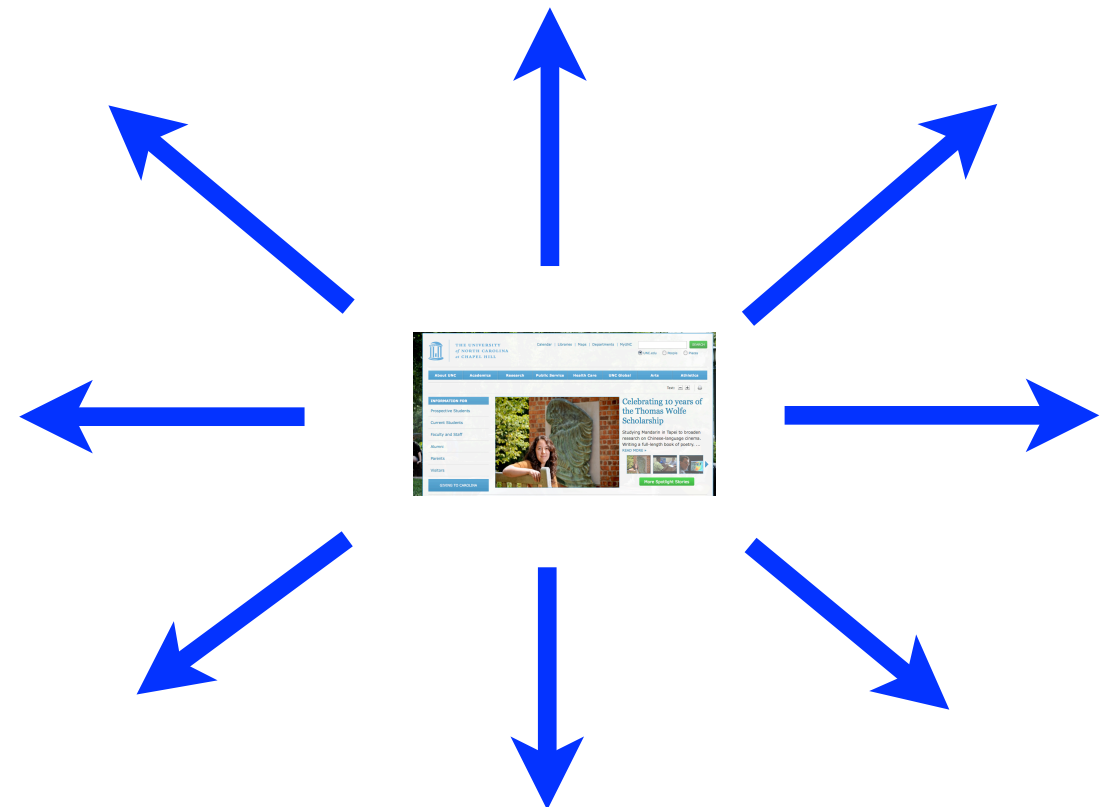
Document Authority

- Number of “endorsements”
 - ▶ **scientific search:** number of citations in other papers
 - ▶ **web search:** number of incoming hyperlinks
 - ▶ **blog search:** number user-generated comments
 - ▶ **twitter search:** number of followers
 - ▶ **review search:** number of times someone found the review useful



Document Authority

- “HUB” score
 - ▶ **scientific search:** number citations of other papers
 - ▶ **web search:** number of outgoing hyperlinks
 - ▶ **blog search:** number of links to other bloggers
 - ▶ **twitter search:** number of people followed by author
 - ▶ **review search:** number of reviews written by the reviewer



Document Prior

$$P(D|Q) \propto P(Q|D) \times P(D)$$

- What is it?
- Anything that affects the likelihood that a document is relevant to any query
 - ▶ document popularity
 - ▶ document authority
 - ▶ amount of content (e.g., length)
 - ▶ topical cohesion
 - ▶ really, you decide ...

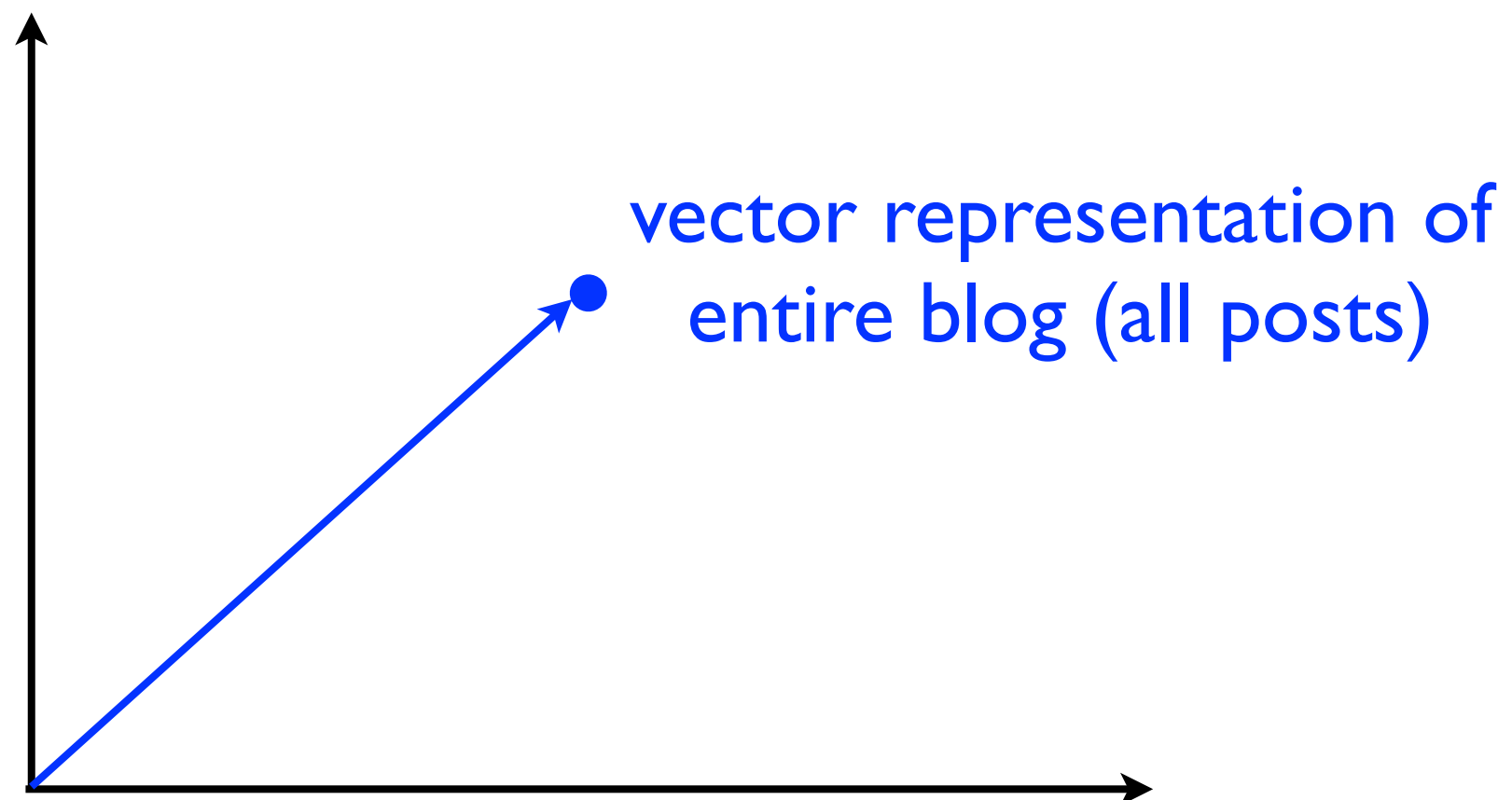
Topical Focus

- **Example:** blog retrieval
- **Objective:** favor blogs that focus on a coherent, recurring topic
- How might we do this? (HINT: vector space model)



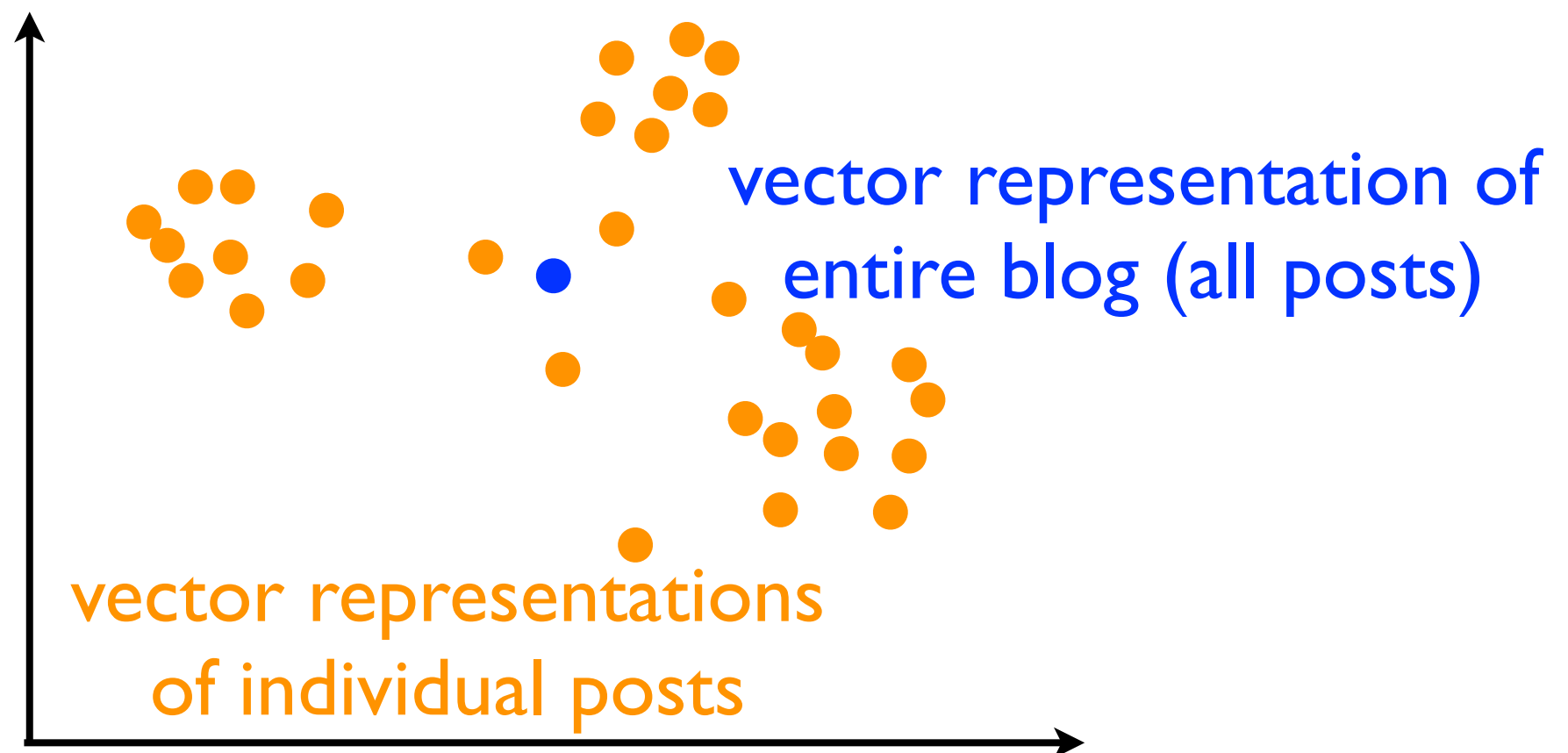
Topical Focus

- **Example:** blog retrieval
- **Objective:** favor blogs that focus on a coherent, recurring topic
- How might we do this? (HINT: vector space model)



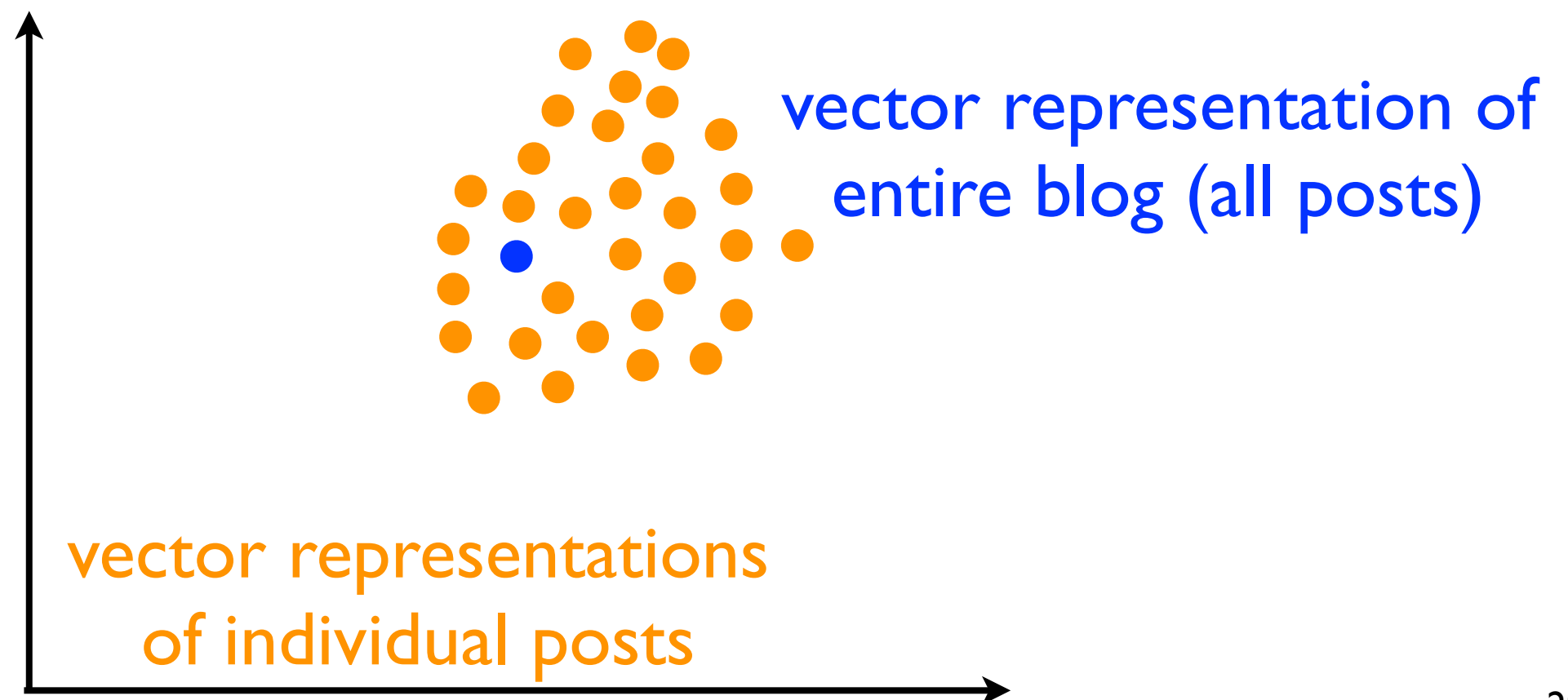
Topical Focus

- How might we do this? (HINT: vector space model)
- Compute average cosine similarity between the **posts** and the entire **blog**



Topical Focus

- How might we do this? (HINT: vector space model)
- Compute average cosine similarity between the **posts** and the entire **blog**



Document Prior

$$P(D|Q) \propto P(Q|D) \times P(D)$$

- What is it?
- Anything you want.
 - ▶ document popularity
 - ▶ document authority
 - ▶ amount of content (e.g., length)
 - ▶ topical focus
 - ▶ really, you decide

What document priors would you use?



Remember Smoothing?

- **YOU:** Are there mountain lions around here?
- **YOUR FRIEND:** Nope.
- **YOU:** How can you be so sure?
- **YOUR FRIEND:** Because I've been hiking here five times before and have never seen one.
- **MOUNTAIN LION:** You should have learned about smoothing by taking INLS 509. Yum!



Remember Smoothing?

- When estimating probabilities, we tend to ...
 - ▶ Over-estimate the probability of observed outcomes
 - ▶ Under-estimate the probability of unobserved outcomes
- The goal of smoothing is to ...
 - ▶ Decrease the probability of observed outcomes
 - ▶ Increase the probability of unobserved outcomes
- Smoothing $P(D)$ is very important!

Example: Click-Rate

$\frac{\text{\# of clicks on the document}}{\text{\# of clicks on any document}}$


$$P(D|Q) \propto P(Q|D) \times P(D)$$

- Do we really want to always give documents that have never been clicked a score of zero?
- How could we smooth this probability?

Example: Click-Rate

$\frac{\text{\# of clicks on the document}}{\text{\# of clicks on any document}}$


$$P(D|Q) \propto P(Q|D) \times P(D)$$

- Do we really want to always give documents that have never been clicked a score of zero?
- Add-one smoothing!

$(\text{\# of clicks on the document}) + 1$

$(\text{\# of clicks on any document}) + (\text{\# of documents})$

Outline

Introduction to language modeling

Language modeling for information retrieval

Query-likelihood Retrieval Model

Smoothing

Priors