

# COEN 169

---

## Document Prior

Yi Fang

Department of Computer Engineering

Santa Clara University

# Outline

---

Introduction to language modeling

Language modeling for information retrieval

Query-likelihood Retrieval Model

Smoothing

Document priors



# Bayes' Law

A photograph of a chalkboard with the formula for Bayes' Law written in blue chalk. The formula is  $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ . The chalkboard is dark, and the blue chalk is clearly visible. The formula is written in a slightly informal, handwritten style.

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

(source: wikipedia)

# Bayes' Law Applied to Ranking

---

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

If we use this formula for ranking,  
which probability does not 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

---

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

- 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

---

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

- 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:

# of clicks on the document

---

# of clicks on any document

# Document Popularity

## most clicked urls - AOL query-log (2006)

---

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

# Document Popularity

## least clicked urls – AOL query-log (2006)

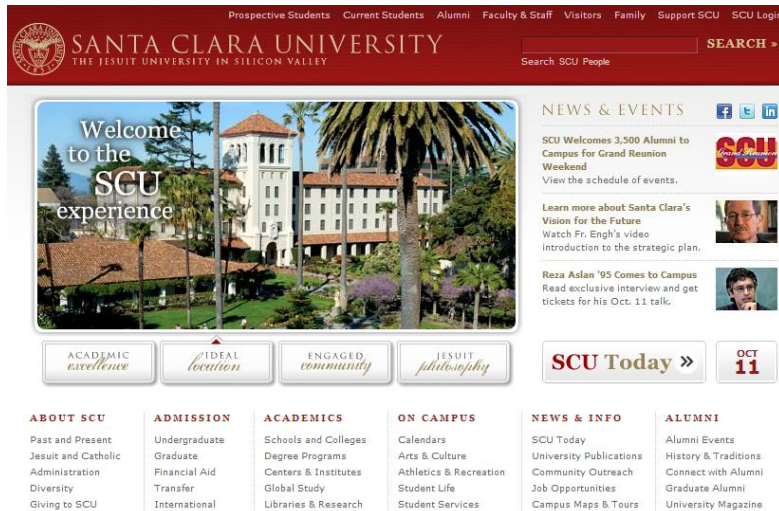
---

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

# Document Popularity

<http://www.scu.edu/>

<http://www.scu.edu/scunews/scutoday/news-views.cfm?c=13215>



- 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

---

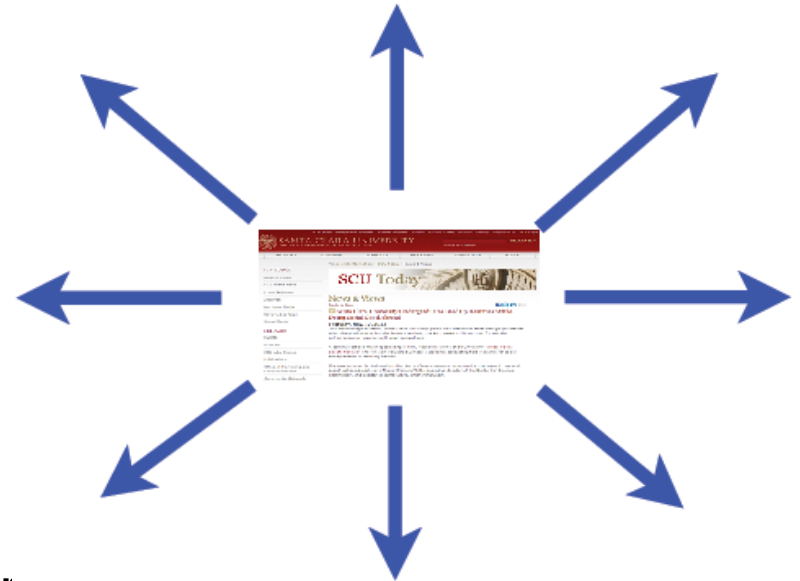
- Number of “endorsements”
  - **scientific search:**  
number of citations in other papers
  - **web search:**  
number of incoming hyperlinks
  - **blog search:**  
number user-generated comments
  - **twitter search:**  
number of followers
  - **review search:**  
number of times someone found the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



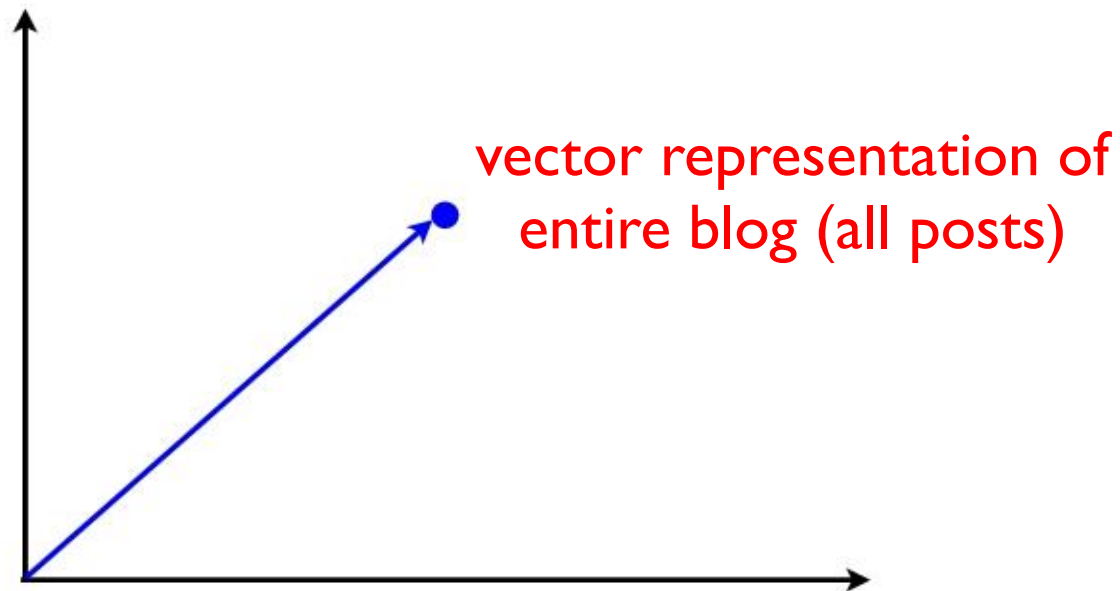
# Topical Focus

---

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

# Topical Focus

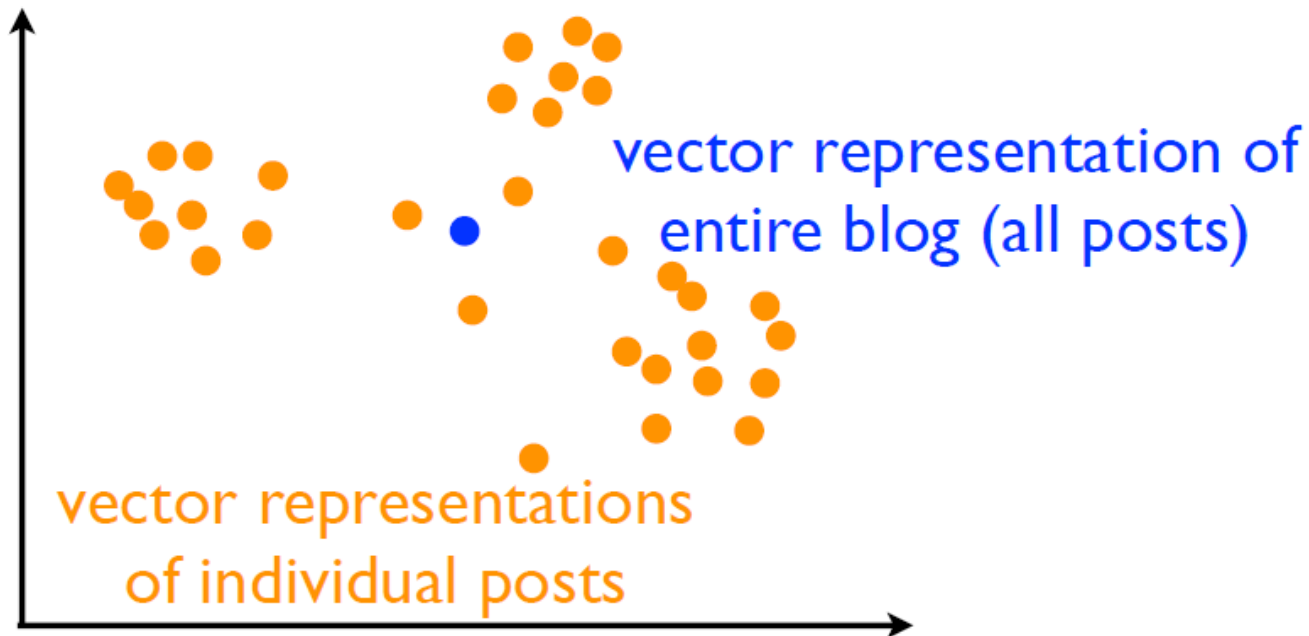
- **Example:** blog retrieval
- **Objective:** favor blogs that focus on a coherent, recurring topic
- How might we do this? (HINT: vector space model)



# Topical Focus

---

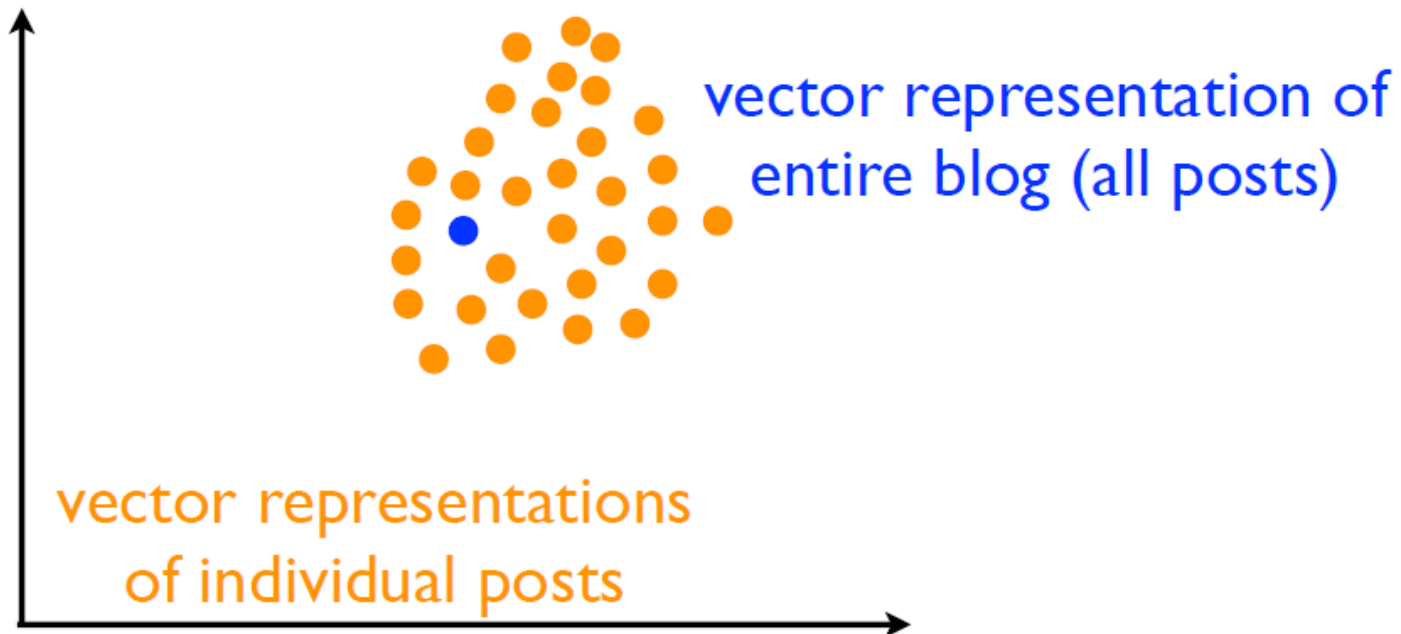
- How might we do this? (HINT: vector space model)
- Compute average cosine similarity between the **posts** and the entire **blog**



# 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

# 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!

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

---

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

# Review

---

Introduction to language modeling

Language modeling for information retrieval

Query-likelihood Retrieval Model

Smoothing

Priors