

# Introduction to **Information Retrieval**

BM25, BM25F, and User Behavior  
Chris Manning and Pandu Nayak

# Summary - BIM

---

- Boils down to

$$RSV^{BIM} = \sum_{x_i=q_i=1} c_i^{BIM}; \quad c_i^{BIM} = \log \frac{p_i(1-r_i)}{(1-p_i)r_i} \quad \text{Log odds ratio}$$

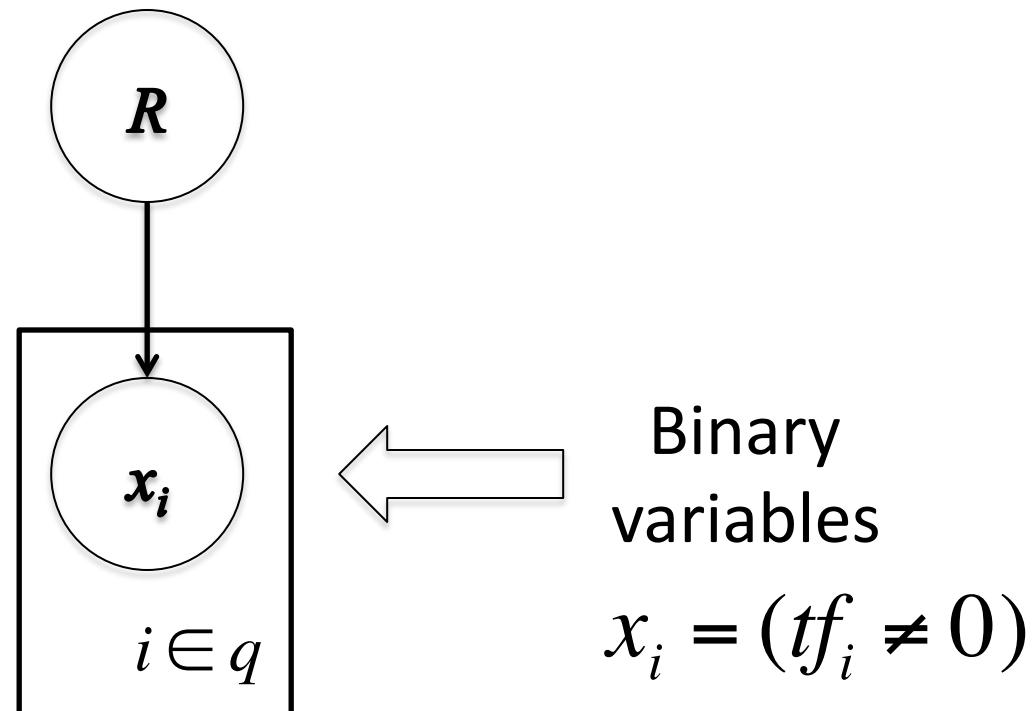
where

	document	relevant (R=1)	not relevant (R=0)
term present	$x_i = 1$	$p_i$	$r_i$
term absent	$x_i = 0$	$(1 - p_i)$	$(1 - r_i)$

- Simplifies to (with constant  $p_i = 0.5$ )

$$RSV = \sum_{x_i=q_i=1} \log \frac{N}{n_i}$$

# Graphical model for BIM



# A key limitation of the BIM

---

- BIM – like much of original IR – was designed for titles or abstracts, and not for modern full text search
- We want to pay attention to term frequency and document lengths, just like in other models we discuss
- Want

$$c_i = \log \frac{p_{tf} r_0}{p_0 r_{tf}}$$

- Want some model of how often terms occur in docs

# 1. Okapi BM25

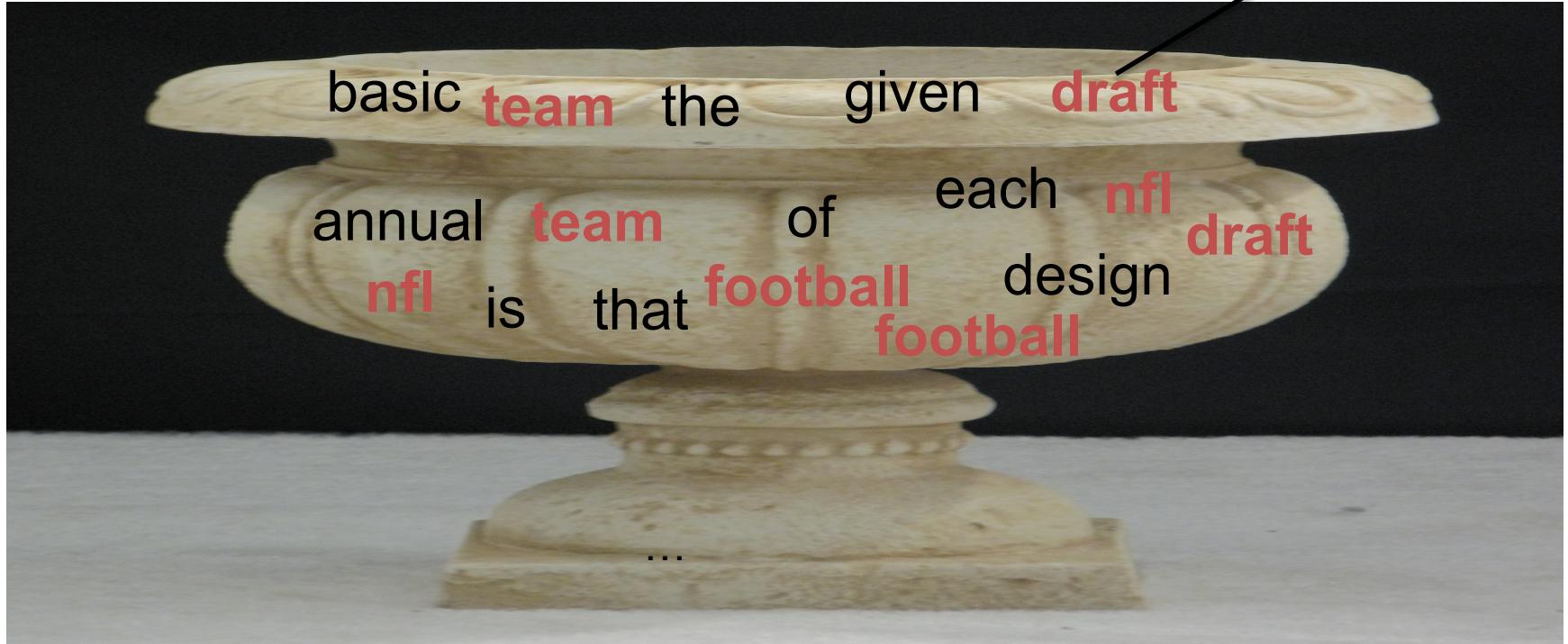
---

- BM25 “Best Match 25” (they had a bunch of tries!)
  - Developed in the context of the Okapi system
  - Started to be increasingly adopted by other teams during the TREC competitions
  - It works well
- Goal: be sensitive to these quantities while not adding too many parameters
  - (Robertson and Zaragoza 2009; Spärck Jones et al. 2000)

# Generative model for documents

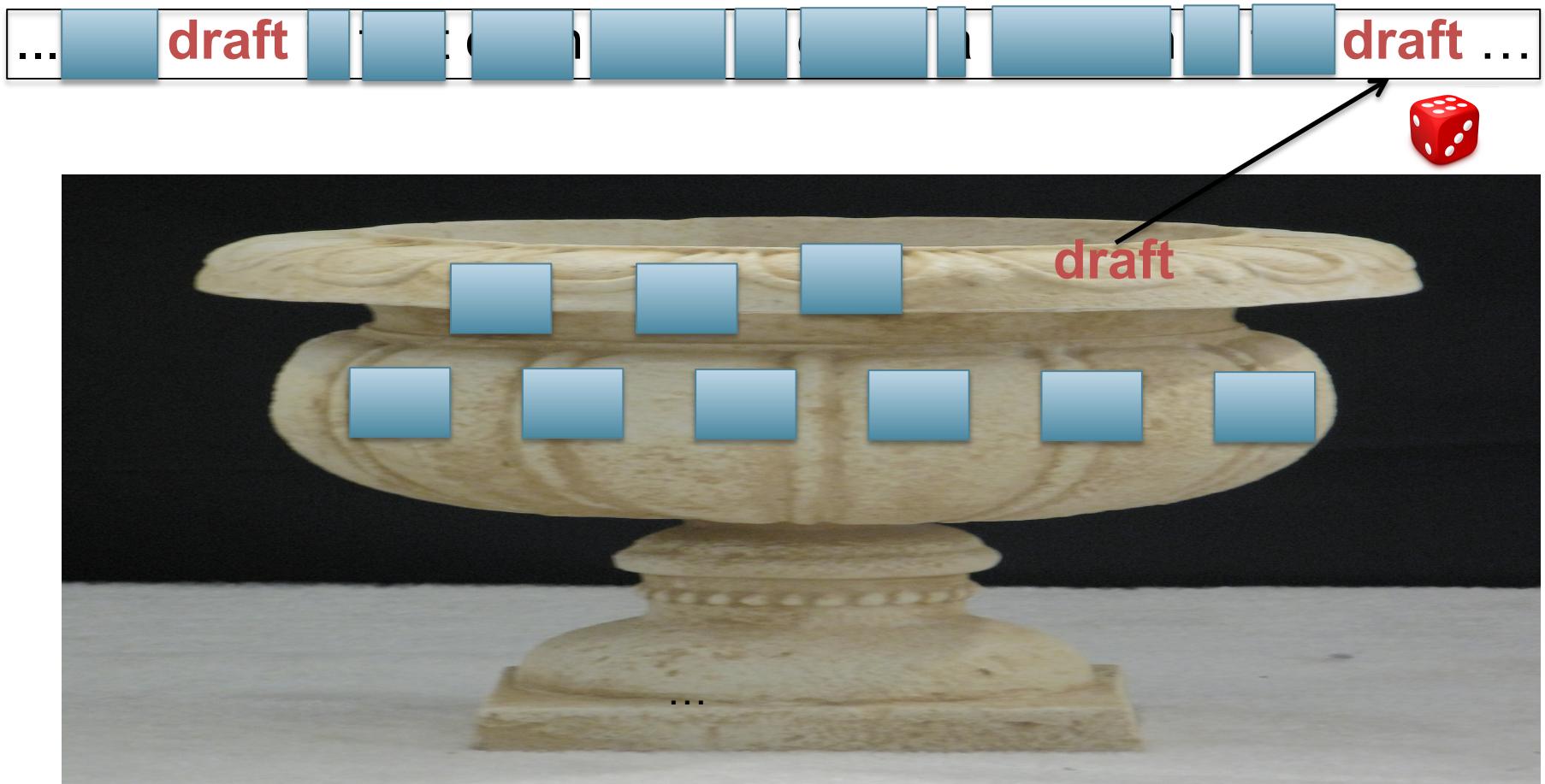
- Words are drawn independently from the vocabulary using a multinomial distribution

... the **draft** is that each **team** is given a position in the **draft** ...



# Generative model for documents

- Distribution of term frequencies ( $tf$ ) follows a binomial distribution – approximates to a Poisson



# Poisson distribution

---

- The Poisson distribution models the probability of  $k$ , the number of events occurring in a fixed interval of time/space, with known average rate  $\lambda$  ( $=cf/T$ ), independent of the last event

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

- Examples
  - Number of cars arriving at the toll booth per minute
  - Number of typos on a page

# Poisson distribution

---

- If  $T$  is large and  $p$  is small, we can approximate a binomial distribution with a Poisson where  $\lambda = Tp$

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

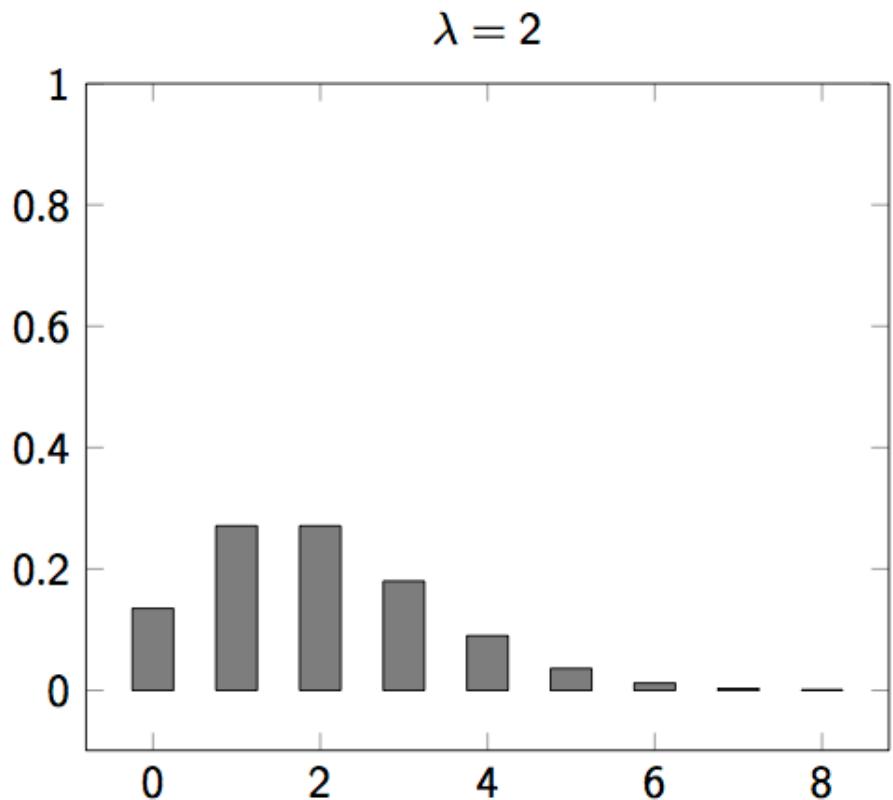
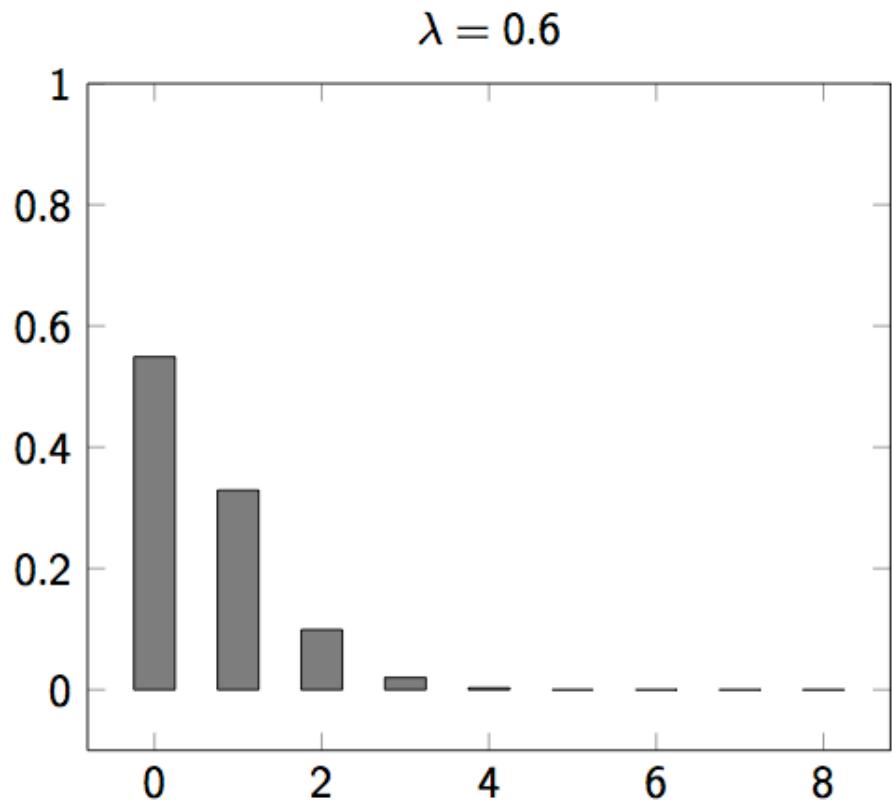
- Mean = Variance =  $\lambda = Tp$ .
- Example  $p = 0.08$ ,  $T = 20$ . Chance of 1 occurrence is:
  - Binomial  $P(1) = \binom{20}{1} (.08)^1 (.92)^{19} = .3282$
  - Poisson  $P(1) = \frac{[(20)(.08)]^1}{1!} e^{-(20)(.08)} = \frac{1.6}{1} e^{-1.6} = 0.3230 \quad \dots \text{already close}$

# Poisson model

---

- Assume that term frequencies in a document ( $tf_i$ ) follow a Poisson distribution
  - “Fixed interval” implies fixed document length ... think roughly constant-sized document abstracts
    - ... will fix later

# Poisson distributions



# (One) Poisson Model

---

- Is a reasonable fit for “general” words
- Is a poor fit for topic-specific words
  - get higher  $p(k)$  than predicted too often

		Documents containing $k$ occurrences of word ( $\lambda = 53/650$ )												
Freq	Word	0	1	2	3	4	5	6	7	8	9	10	11	12
53	expected	599	49	2										
52	based	600	48	2										
53	conditions	604	39	7										
55	cathexis	619	22	3	2	1	2	0	1					
51	comic	642	3	0	1	0	0	0	0	0	0	1	1	2

Harter, “A Probabilistic Approach to Automatic Keyword Indexing”, JASIST, 1975

# Eliteness (“aboutness”)

---

- Model term frequencies using *eliteness*
- What is eliteness?
  - Hidden variable for each document-term pair, denoted as  $E_i$  for term  $i$
  - Represents *aboutness*: a term is elite in a document if, in some sense, the document is about the concept denoted by the term
  - Eliteness is binary
  - Term occurrences depend only on eliteness...
  - ... but eliteness depends on relevance

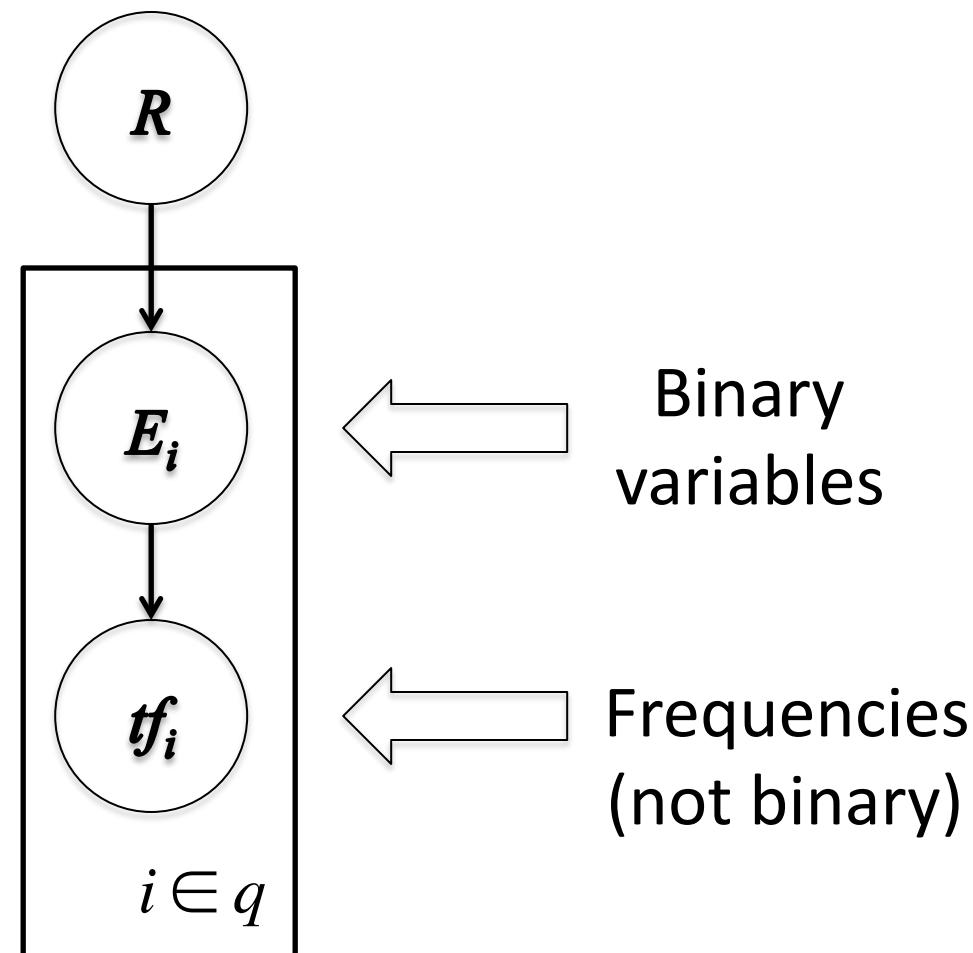
# Elite terms

---

Text from the Wikipedia page on the NFL draft showing  
**elite terms**

The **National Football League Draft** is an annual event in which the **National Football League (NFL)** teams select eligible college football players. It serves as the league's most common source of player recruitment. The basic design of the **draft** is that each **team** is given a **position** in the **draft order** in **reverse order** relative to its **record** ...

# Graphical model with eliteness



# Retrieval Status Value

---

- Similar to the BIM derivation, we have

$$RSV^{elite} = \sum_{i \in q, tf_i > 0} c_i^{elite}(tf_i);$$

where

$$c_i^{elite}(tf_i) = \log \frac{p(TF_i = tf_i | R = 1)p(TF_i = 0 | R = 0)}{p(TF_i = 0 | R = 1)p(TF_i = tf_i | R = 0)}$$

and using eliteness, we have:

$$\begin{aligned} p(TF_i = tf_i | R) &= p(TF_i = tf_i | E_i = elite)p(E_i = elite | R) \\ &\quad + p(TF_i = tf_i | E_i = \overline{elite})(1 - p(E_i = elite | R)) \end{aligned}$$

## 2-Poisson model

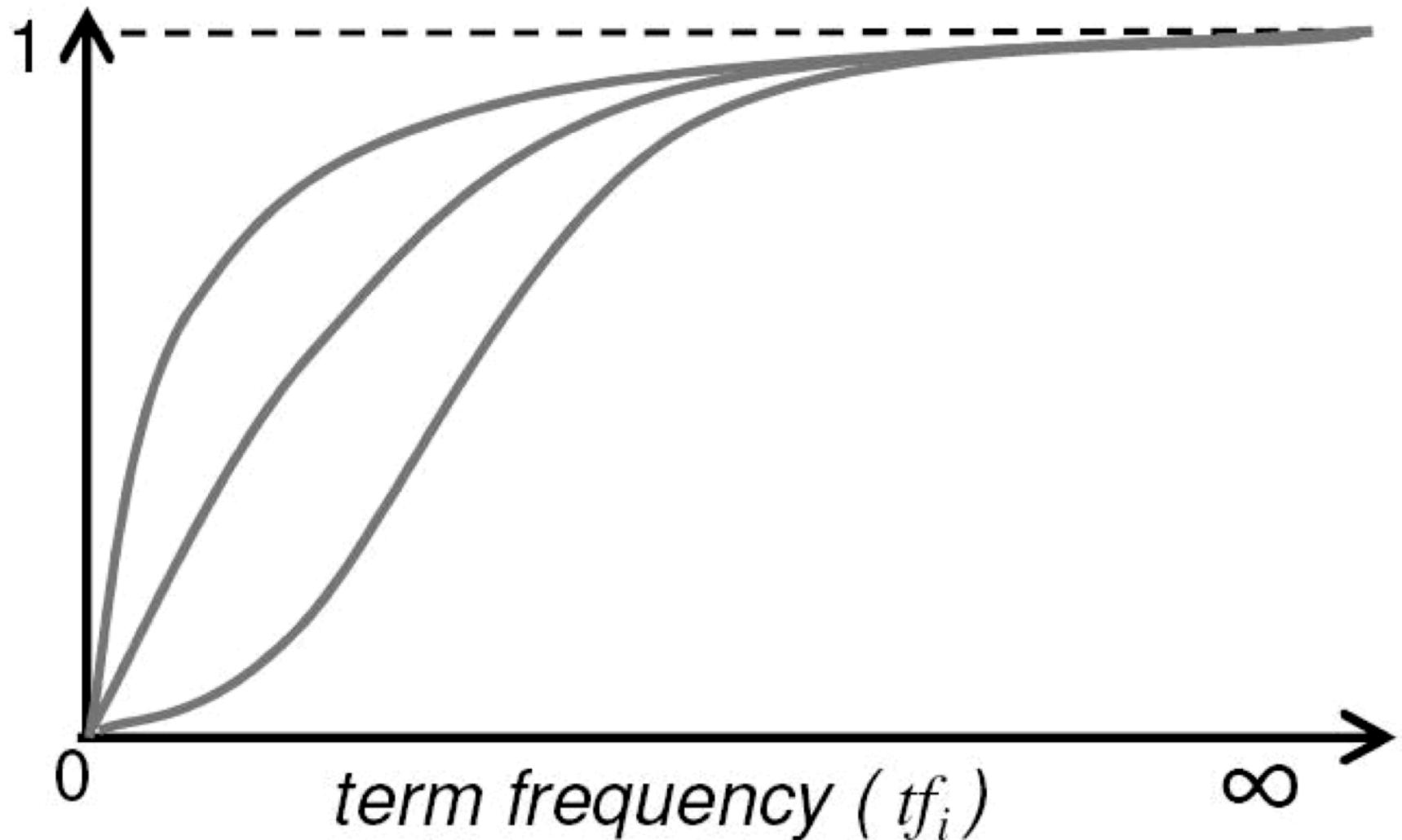
---

- The problems with the 1-Poisson model suggests fitting two Poisson distributions
- In the “2-Poisson”, the distribution is different depending on whether the term is elite or not

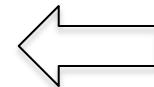
$$p(TF_i = k_i | R) = \pi \frac{\lambda^k}{k!} e^{-\lambda} + (1 - \pi) \frac{\mu^k}{k!} e^{-\mu}$$

- where  $\pi$  is probability that document is elite for term
- but, unfortunately, we don't know  $\pi, \lambda, \mu$

Let's get an idea: Graphing  $c_i^{elite}(tf_i)$  for different parameter values of the 2-Poisson



# Qualitative properties

- $c_i^{elite}(0) = 0$
- $c_i^{elite}(tf_i)$  increases monotonically with  $tf_i$
- ... but asymptotically approaches a maximum value as  $tf_i \rightarrow \infty$  [not true for simple scaling of tf]
- ... with the asymptotic limit being  $c_i^{BIM}$   Weight of  
eliteness  
feature

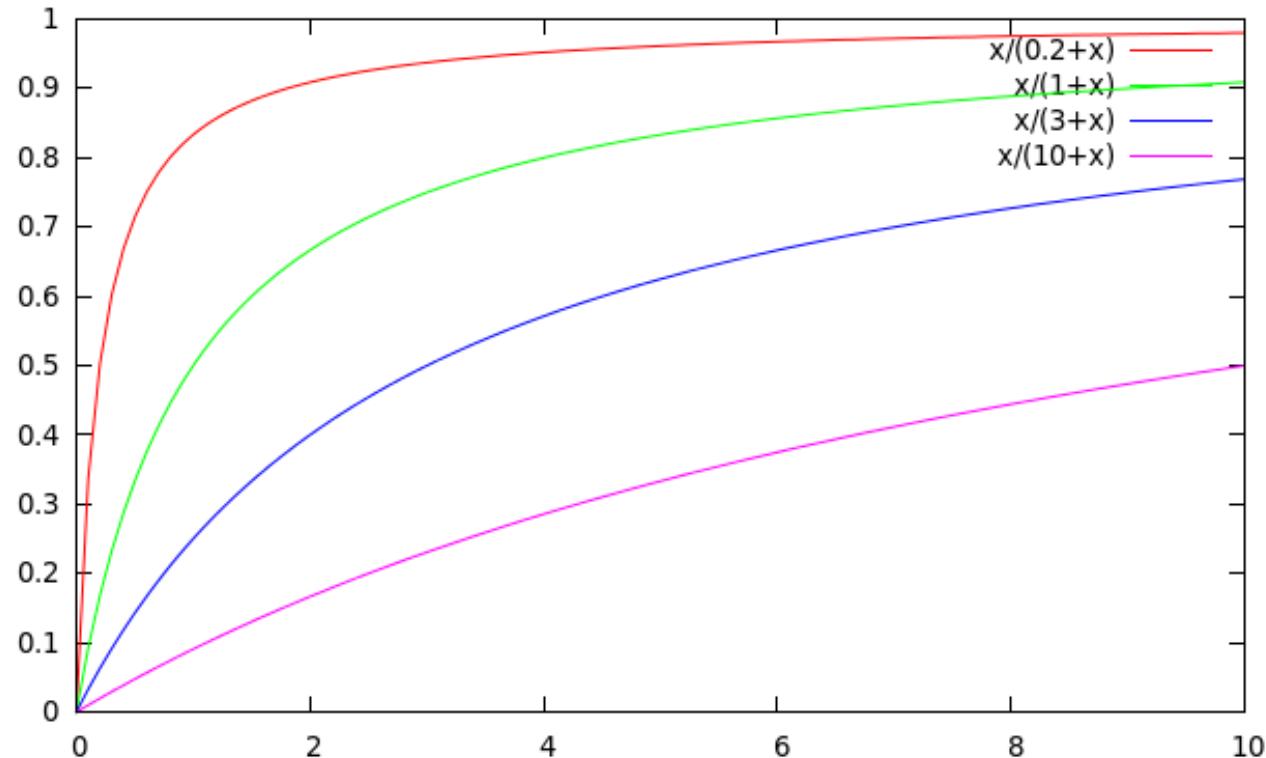
# Approximating the saturation function

---

- Estimating parameters for the 2-Poisson model is not easy
- ... So approximate it with a simple parametric curve that has the same qualitative properties

$$\frac{tf}{k_1 + tf}$$

# Saturation function



- For high values of  $k_1$ , increments in  $tf_i$  continue to contribute significantly to the score
- Contributions tail off quickly for low values of  $k_1$

# “Early” versions of BM25

- Version 1: using the saturation function

$$c_i^{BM25v1}(tf_i) = c_i^{BIM} \frac{tf_i}{k_1 + tf_i}$$

- Version 2: BIM simplification to IDF

$$c_i^{BM25v2}(tf_i) = \log \frac{N}{df_i} \times \frac{(k_1 + 1)tf_i}{k_1 + tf_i}$$

- $(k_1 + 1)$  factor doesn't change ranking, but makes term score 1 when  $tf_i = 1$
- Similar to  $tf-idf$ , but term scores are bounded

# Document length normalization

---

- Longer documents are likely to have larger  $tf_i$  values
- Why might documents be longer?
  - Verbosity: suggests observed  $tf_i$  too high
  - Larger scope: suggests observed  $tf_i$  may be right
- A real document collection probably has both effects
- ... so should apply some kind of normalization

# Document length normalization

---

- Document length:

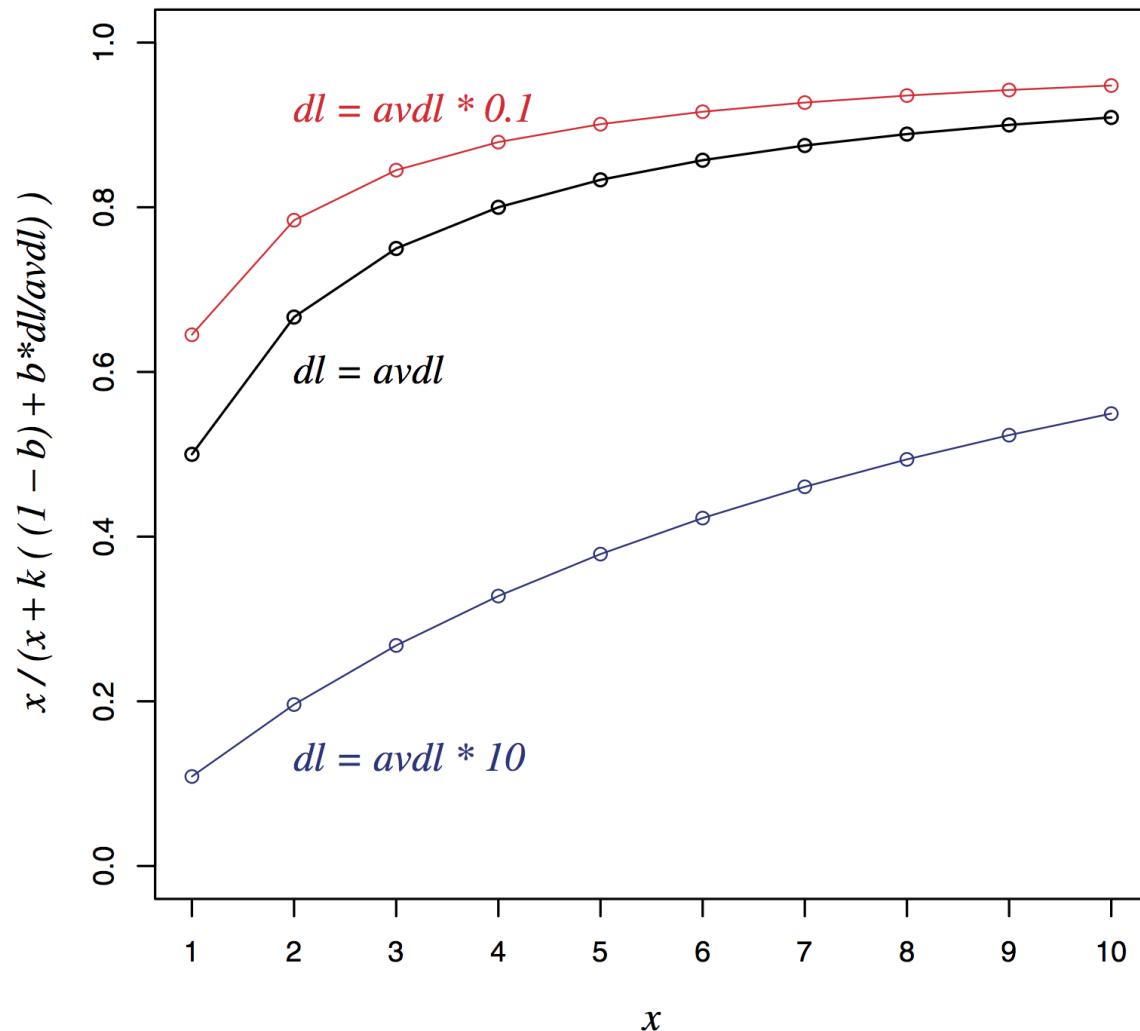
$$dl = \sum_{i \in V} tf_i$$

- $avdl$ : Average document length over collection
- Length normalization component

$$B = \left( (1 - b) + b \frac{dl}{avdl} \right), \quad 0 \leq b \leq 1$$

- $b = 1$  full document length normalization
- $b = 0$  no document length normalization

# Document length normalization



# Okapi BM25

- Normalize  $tf$  using document length

$$tf'_i = \frac{tf_i}{B}$$

$$\begin{aligned} c_i^{BM25}(tf_i) &= \log \frac{N}{df_i} \times \frac{(k_1 + 1)tf'_i}{k_1 + tf'_i} \\ &= \log \frac{N}{df_i} \times \frac{(k_1 + 1)tf_i}{k_1((1 - b) + b \frac{dl}{avdl}) + tf_i} \end{aligned}$$

- BM25 ranking function

$$RSV^{BM25} = \sum_{i \in q} c_i^{BM25}(tf_i);$$

# Okapi BM25

---

$$RSV^{BM25} = \sum_{i \in q} \log \frac{N}{df_i} \cdot \frac{(k_1 + 1)tf_i}{k_1((1 - b) + b \frac{dl}{avdl}) + tf_i}$$

- $k_1$  controls term frequency scaling
  - $k_1 = 0$  is binary model;  $k_1 = 1$  is raw term frequency
- $b$  controls document length normalization
  - $b = 0$  is no length normalization;  $b = 1$  is relative frequency (fully scale by document length)
- Typically,  $k_1$  is set around 1.2–2 and  $b$  around 0.75
- IIR sec. 11.4.3 discusses incorporating query term weighting and (pseudo) relevance feedback

## 2. Ranking with features

---

- Textual features
  - Zones: Title, author, abstract, body, anchors, ...
  - Proximity
  - ...
- Non-textual features
  - File type
  - File age
  - Page rank
  - ...

# Ranking with zones

---

- Straightforward idea:
  - Apply your favorite ranking function (BM25) to each zone separately
  - Combine zone scores using a weighted linear combination
- But that seems to imply that the eliteness properties of different zones are different and independent of each other
  - ...which seems unreasonable

# Ranking with zones

---

- Alternate idea
  - Assume eliteness is a term/document property shared across zones
  - ... but the relationship between eliteness and term frequencies are zone-dependent
    - e.g., denser use of elite topic words in title
- Consequence
  - First combine evidence across zones for each term
  - Then combine evidence across terms

# BM25F with zones

- Calculate a weighted variant of total term frequency
- ... and a weighted variant of document length

$$\tilde{tf}_i = \sum_{z=1}^Z v_z tf_{zi} \quad \tilde{dl} = \sum_{z=1}^Z v_z len_z \quad avd\tilde{l} = \text{Average } \tilde{dl} \text{ across all documents}$$

where

$v_z$  is zone weight

$tf_{zi}$  is term frequency in zone  $z$

$len_z$  is length of zone  $z$

$Z$  is the number of zones

# Simple BM25F with zones

---

$$RSV^{SimpleBM25F} = \sum_{i \in q} \log \frac{N}{df_i} \cdot \frac{(k_1 + 1)\tilde{tf}_i}{k_1((1 - b) + b \frac{\tilde{dl}}{avd\tilde{l}}) + \tilde{tf}_i}$$

- Simple interpretation: zone  $z$  is “replicated”  $v_z$  times
- But we may want zone-specific parameters ( $k_1$ ,  $b$ , IDF)

# BM25F

---

- Empirically, zone-specific length normalization (i.e., zone-specific  $b$ ) has been found to be useful

$$\tilde{tf}_i = \sum_{z=1}^Z v_z \frac{tf_{zi}}{B_z}$$

$$B_z = \left( (1 - b_z) + b_z \frac{len_z}{avlen_z} \right), \quad 0 \leq b_z \leq 1$$

$$RSV^{BM25F} = \sum_{i \in q} \log \frac{N}{df_i} \cdot \frac{(k_1 + 1)\tilde{tf}_i}{k_1 + \tilde{tf}_i}$$

See Robertson and Zaragoza (2009: 364)

# Ranking with non-textual features

---

- Assumptions
  - Usual independence assumption
    - Independent of each other and of the textual features
    - Allows us to factor out  $\frac{p(F_j = f_j | R = 1)}{p(F_j = f_j | R = 0)}$  in BIM-style derivation
  - Relevance information is ***query independent***
    - Usually true for features like page rank, age, type, ...
    - Allows us to keep all non-textual features in the BIM-style derivation where we drop non-query terms

# Ranking with non-textual features

$$RSV = \sum_{i \in q} c_i(tf_i) + \sum_{j=1}^F \lambda_j V_j(f_j)$$

where

$$V_j(f_j) = \log \frac{p(F_j = f_j | R = 1)}{p(F_j = f_j | R = 0)}$$

and  $\lambda_j$  is an artificially added free parameter to account for rescalings in the approximations

- Care must be taken in selecting  $V_j$  depending on  $F_j$ . E.g.

$$\log(\lambda'_i + f_j) \quad \frac{f_j}{\lambda'_i + f_j} \quad \frac{1}{\lambda'_i + \exp(-f_j \lambda''_j)}$$

- Explains why  $RSV^{BM25} + \log(\text{pagerank})$  works well

Taken with slight adaptation from Fan Guo and Chao Liu's 2009 CIKM tutorial: Statistical Models for Web Search: Click Log Analysis

# User Behavior

- Search Results for “CIKM”

Web [Show options...](#) Results 1 - 30 of about 144,000 for cikm. (0.44 seconds)

[CIKM 2009 | Home](#)  
Nov 3, 2009 ... CIKM 2009 (The 18th ACM Conference on Information and Knowledge Management) will be held on November 2-6, 2009, Hong Kong. ...  
[Accepted Papers](#) - [Papers](#) - [Workshops](#) - [Registration](#)  
[www.comp.polyu.edu.hk/conference/cikm2009/](http://www.comp.polyu.edu.hk/conference/cikm2009/) - [Cached](#) - [Similar](#)

[Accepted Papers - CIKM 2009 | Home](#)  
Accepted Papers. Full Research Papers; Short Research Papers; Demonstration Papers. Full Papers Space-Economical Partial Gram Indices for Exact Substring ...  
[www.comp.polyu.edu.hk/.../cikm2009/.../accepted\\_papers.htm](http://www.comp.polyu.edu.hk/.../cikm2009/.../accepted_papers.htm) - [Cached](#) - [Similar](#)

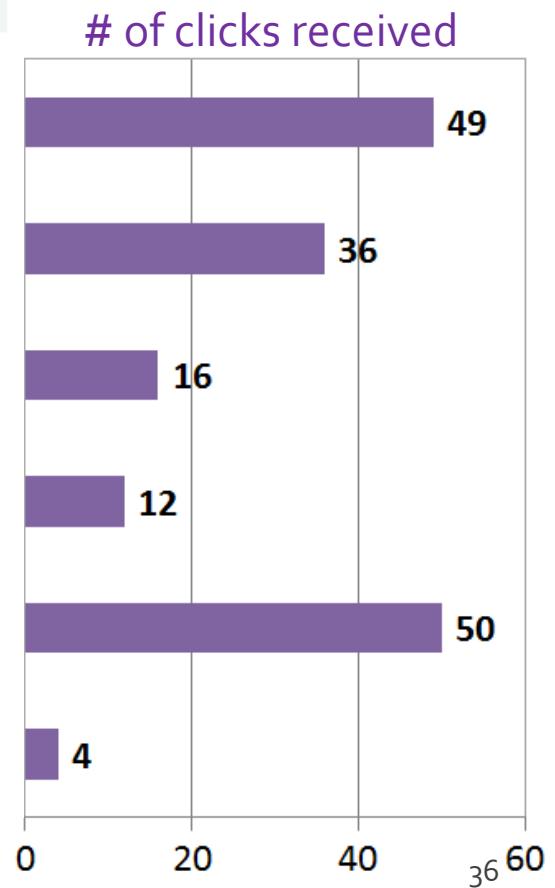
[Show more results from www.comp.polyu.edu.hk](#)

[Conference on Information and Knowledge Management \(CIKM\)](#)  
The Conference on Information and Knowledge Management (CIKM) provides an international forum for presentation and discussion of research on information and ...  
[www.cikmconference.org/](http://www.cikmconference.org/) - [Cached](#)

[CIKM 2008 | Home](#)  
CIKM 2008 will take place at the Napa Valley Marriott Hotel & Spa, 3425 Solano Avenue, Napa, CA 94558. This is right in the center of beautiful wine country ...  
[Papers](http://www.cikm2008.org/) - [Posters](#) - [Student Travel](#) - [Workshops](#)  
[www.cikm2008.org/](http://www.cikm2008.org/) - [Cached](#) - [Similar](#)

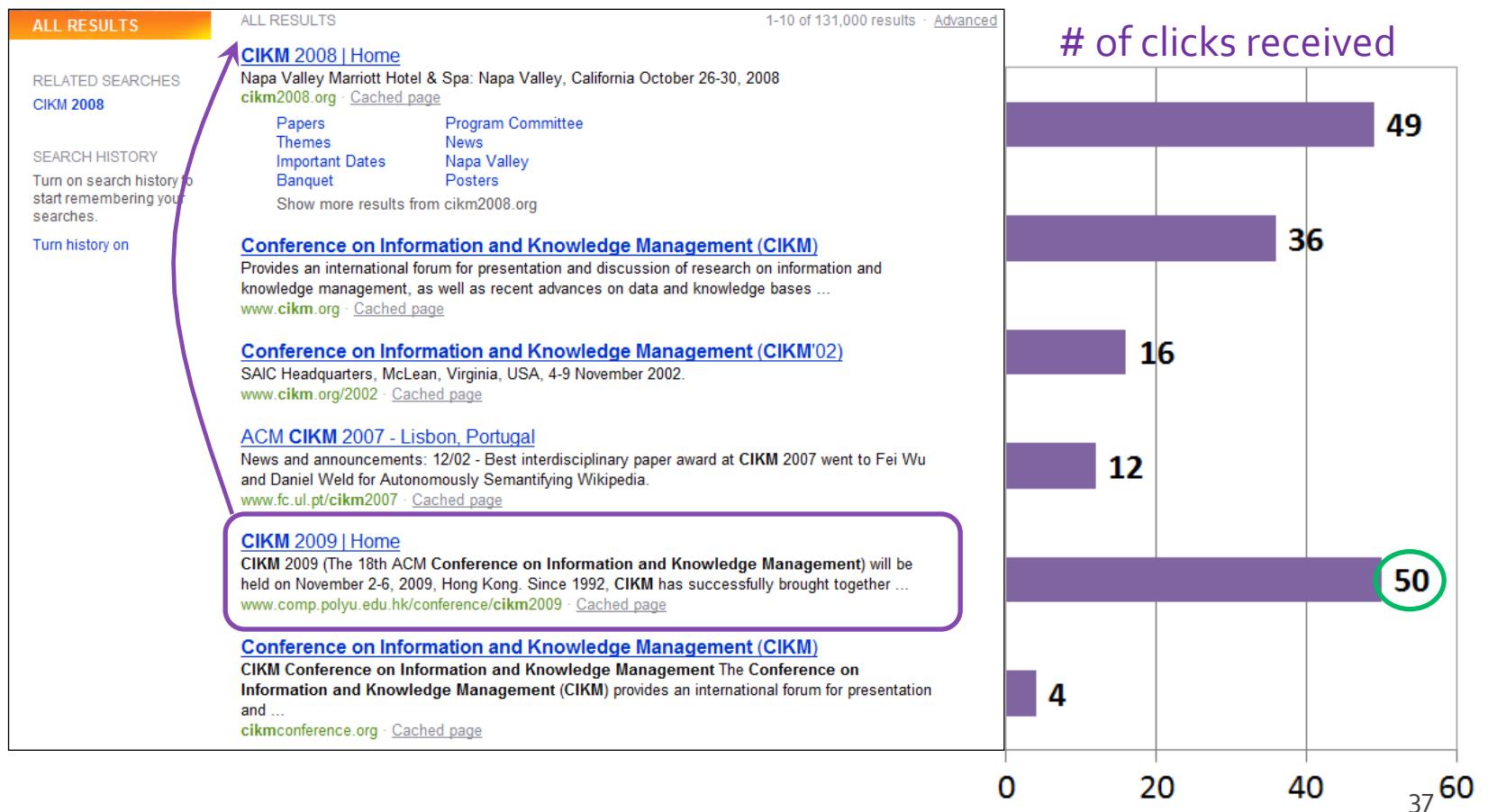
[The 19th ACM International Conference on Information and Knowledge ...](#)  
CIKM 2010 will take place in Toronto, Canada. Toronto is Canada's largest city. More than 60% of the US population live within a 90-minute flight to Toronto ...  
[www.yorku.ca/cikm10/](http://www.yorku.ca/cikm10/) - [Cached](#)

[CIKM 2010 : International Conference on Information and Knowledge ...](#)  
CIKM 2010 : International Conference on Information and Knowledge Management - Conference and Journal.  
[www.wikicfp.com/cfp/servlet/event.showcfp?eventid...2](http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid...2) - [Cached](#) - [Similar](#)



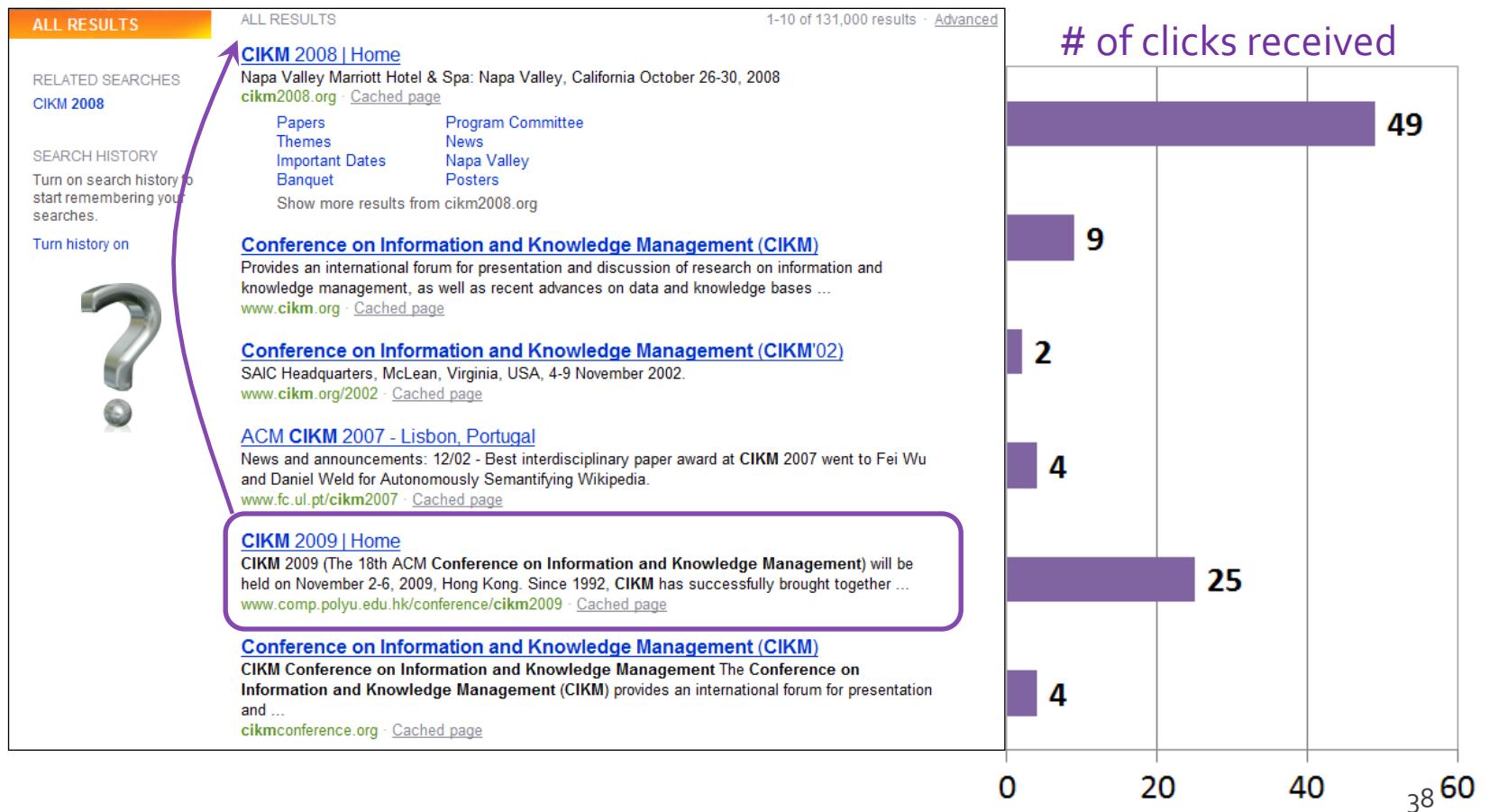
# User Behavior

- Adapt ranking to user clicks?



# User Behavior

- Tools needed for non-trivial cases



# Web search click log

An example

user-ct-test-collection-06.txt					
1998497 anthony burger	2006-03-05 13:01:36	2	http://www.anthonyburger.com		
1998497 gaither	2006-03-05 13:02:22	4	http://www.bill.gaither.com-music.homepages.org		
1998497 allegiant air	2006-03-05 15:27:59	1	http://www.allegiantair.com		
1998497 gaitho	2006-03-05 17:07:32				
1998497 gaither	2006-03-05 17:07:44	7	http://www.gaither.com		
1998497 gaitho	2006-03-05 17:09:53				
1998497 gaither	2006-03-05 17:10:03	7	http://www.gaither.com		
1998497 allegiant air	2006-03-05 18:22:26	1	http://www.allegiantair.com		
1998497 disney coronado springs resort orlando fl		2006-03-07 14:09:00	5	http://hotels.about.com	
1998497 www.hli.com	2006-03-10 09:05:39				
1998497 heritage lottery international	2006-03-10 09:06:56	1	http://blog.supersurge.com		
1998497 googlemaps.com	2006-03-11 00:12:28	1	http://www.googlemaps.com		
1998497 amy grant	2006-03-11 19:29:34	7	http://www.mindspring.com		
1998497 amy grant	2006-03-11 19:29:34	2	http://www.amygrant.com		
1998497 amy grant	2006-03-11 19:29:34	5	http://en.wikipedia.org		
1998497 david phelps	2006-03-11 19:33:55	1	http://www.davidphelps.com		
1998497 imercer.com social security	2006-03-12 13:58:18				
1998497 imercer.com social security	2006-03-12 13:58:30				
1998497 www.uhc.com	2006-03-12 15:07:01	1	http://www.uhc.com		
1998497 www.metlife.com	2006-03-12 15:31:06	2	http://www.metlife.com		
1998497 www.vsp.com	2006-03-12 15:36:37	1	http://www.vsp.com		
1998497 www.birdsandblooms.com	2006-03-15 20:06:15				
1998497 www.birdsandblooms.com	2006-03-15 20:06:27	2	http://www.birdsandblooms.com		
1998497 yahoo.com	2006-03-18 13:32:15	1	http://www.yahoo.com		
1998497 google.com	2006-03-18 13:51:35	1	http://www.google.com		
1998497 google.com	2006-03-18 14:13:57				
1998497 google.com	2006-03-18 14:14:25				
1998497 google.com	2006-03-18 14:14:52				
1998497 google.com	2006-03-18 14:15:17				
1998497 google.com	2006-03-18 14:15:54				
1998497 google.com people	2006-03-18 14:16:17				
1998497 www.bostonmarket.com	2006-03-20 19:48:30	1	http://www.bostonmarket.com		
1998497 american heart association	2006-03-24 16:58:34	1	http://www.americanheart.org		
1998497 american cancer society	2006-03-24 19:45:55	5	http://www.acs-tx.org		

# Web Search Click Log

- How large is the click log?

- **bing** search logs: 10+ TB/day

- In existing publications:

- [Craswell+08]: 108k sessions
    - [Dupret+08] : 4.5M sessions (21 subsets \* 216k sessions)
    - [Guo +09a] : 8.8M sessions from 110k unique queries
    - [Guo+09b]: 8.8M sessions from 110k unique queries
    - [Chapelle+09]: 58M sessions from 682k unique queries
    - [Liu+09a]: 0.26PB data from 103M unique queries



# Interpret Clicks: an Example

[CIKM 2008 | Home](#) Napa Valley Marriott Hotel pa: Napa Valley, California October 26-30, 2008  
[cikm2008.org](#) · [Cached page](#)

Papers Program Committee  
Themes News  
Important Dates Napa Valley  
Banquet Posters

Show more results from cikm2008.org

**Conference on Information and Knowledge Management (CIKM)**  
Provides an international forum for presentation and discussion of research on information and knowledge management, as well as recent advances on data and knowledge bases ...  
[www.cikm.org](#) · [Cached page](#)

**Conference on Information and Knowledge Management (CIKM'02)**  
SAIC Headquarters, McLean, Virginia, USA, 4-9 November 2002.  
[www.cikm.org/2002](#) · [Cached page](#)

**ACM CIKM 2007 - Lisbon, Portugal**  
News and announcements: 12/02 - Best interdisciplinary paper award at CIKM 2007 went to Fei Wu and Daniel Weld for Autonomously Semantifying Wikipedia.  
[www.fc.ul.pt/cikm2007](#) · [Cached page](#)

[CIKM 2009 | Home](#) CIKM 2009 (The 18th ACM Conference on Information and Knowledge Management) will be held on November 2-6, 2009, Hong Kong. Since 1992, CIKM has successfully brought together ...  
[www.comp.polyu.edu.hk/conference/cikm2009](#) · [Cached page](#)

**Conference on Information and Knowledge Management (CIKM)**  
CIKM Conference on Information and Knowledge Management The Conference on Information and Knowledge Management (CIKM) provides an international forum for presentation and ...  
[cikmconference.org](#) · [Cached page](#)

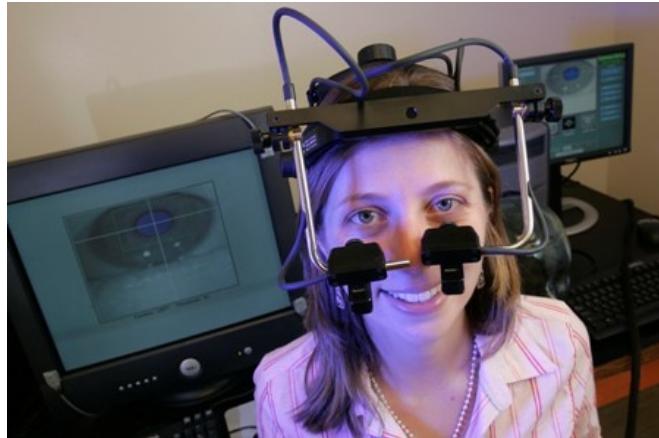
**CIKM 2004**  
Identify challenging problems facing the development of future knowledge and information systems, and shape future directions of research by soliciting and reviewing high quality ...  
[ir.iit.edu/cikm2004](#) · [Cached page](#)

**CIKM**  
International Conference on Information and Knowledge Management (CIKM) CIKM Home Page  
ACM DL: CIKM 17. CIKM 2008: Napa Valley, California, USA. James G. Shanahan, Sihem Amer-Yahia ...  
[www.informatik.uni-trier.de/~ley/db/conf/cikm/index.html](#) · [Cached page](#)

- Clicks are good...
  - Are these two clicks equally “good”?
- Non-clicks may have excuses:
  - Not relevant
  - Not examined



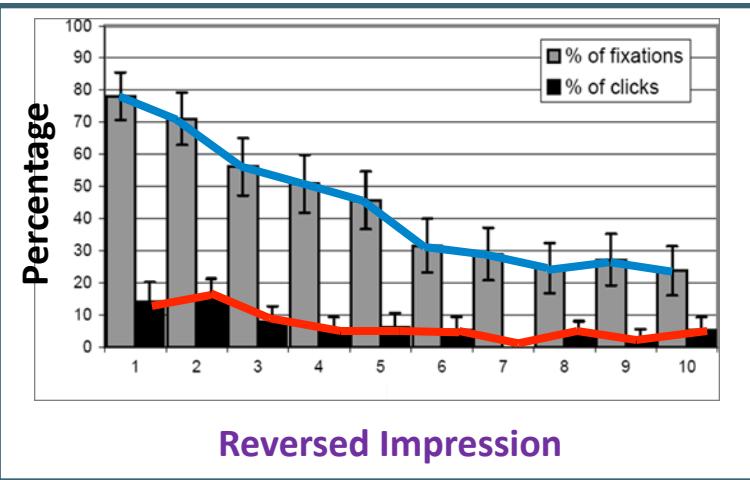
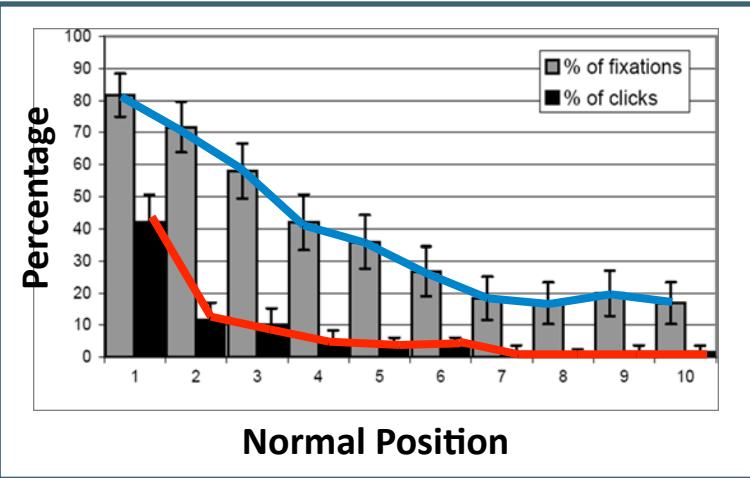
# Eye-tracking User Study



A screenshot of a Bing search results page. The search term 'cikm' is entered in the search bar. The results show 1-10 of 131,000 results. The first result is 'CIKM 2008 | Home' from Napa Valley Marriott Hotel &amp; Spa, Napa Valley, California, October 26-30, 2008. The second result is 'Conference on Information and Knowledge Management (CIKM)' from SAI. The third result is 'ACM CIKM 2007 - Lisbon, Portugal'. The fourth result is 'CIKM 2009 - The 18th ACM Conference on Information and Knowledge Management'. The fifth result is 'Conference on Information and Knowledge Management (CIKM)' from CIKM Conference on Information and Knowledge Management. The sixth result is 'CIKM 2004'.



# Click Position-bias



- Higher positions receive more user attention (eye fixation) and clicks than lower positions.
- This is true even in the extreme setting where the order of positions is reversed.
- “Clicks are informative but biased”. [Joachims+07]

# User behavior

---

- User behavior is an intriguing source of relevance data
  - Users make (somewhat) informed choices when they interact with search engines
  - Potentially a lot of data available in search logs
- But there are significant caveats
  - User behavior data can be very noisy
  - Interpreting user behavior can be tricky
  - Spam can be a significant problem
  - Not all queries will have user behavior

# Features based on user behavior

---

From [Agichtein, Brill, Dumais 2006; Joachims 2002]

- Click-through features
  - Click frequency, click probability, click deviation
  - Click on next result? previous result? above? below>?
- Browsing features
  - Cumulative and average time on page, on domain, on URL prefix; deviation from average times
  - Browse path features
- Query-text features
  - Query overlap with title, snippet, URL, domain, next query
  - Query length

# Incorporating user behavior into ranking algorithm

---

- Incorporate user behavior features into a ranking function like BM25F
  - But requires an understanding of user behavior features so that appropriate  $V_j$  functions are used
- Incorporate user behavior features into *learned* ranking function
- Either of these ways of incorporating user behavior signals improve ranking

# Resources

---

- S. E. Robertson and H. Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* 3(4): 333-389.
- K. Spärck Jones, S. Walker, and S. E. Robertson. 2000. A probabilistic model of information retrieval: Development and comparative experiments. Part 1. *Information Processing and Management* 779–808.
- T. Joachims. Optimizing Search Engines using Clickthrough Data. 2002. *SIGKDD*.
- E. Agichtein, E. Brill, S. Dumais. 2006. Improving Web Search Ranking By Incorporating User Behavior Information. 2006. *SIGIR*.