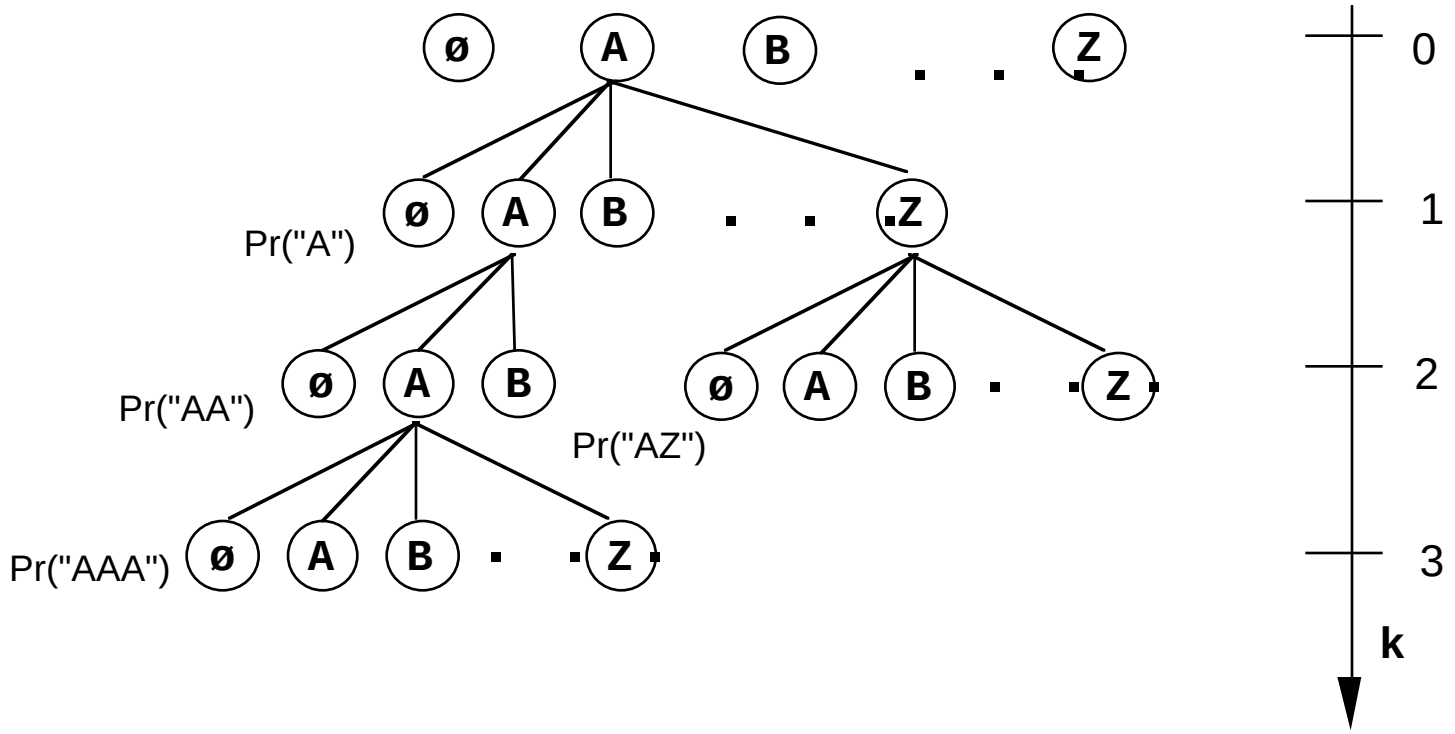
Lexicographic Tree



- Fig 5.1

Length distribution of words

$$N_k \equiv \text{number words of length } \leq k$$

$$= \sum_{i=1}^{k} M^i = \frac{M(1-M^k)}{1-M}$$

- via standard identity

Word: bracketed by spaces

$$p_k \equiv \text{word of length } k$$

$$= p^{k+2}$$

$$\propto \frac{1}{(M+1)^{k+2}}$$

$$= \frac{c}{(M+1)^{k+2}}$$

Constant of proportionality

$$\sum_{k=1}^{\infty} c \cdot M^k p_k = 1$$

$$c = \frac{(M+1)^2}{M}$$

•

Probability implies rank

$$N_{k-1} < r_k \leq N_k$$

$$\tilde{r} = \frac{N_{k-1} + 1 + N_k}{2}$$

$$= (M^k - 1)\frac{M+1}{2(M-1)}$$

- Both probability and rank functions of length
- Solve both for for k (length)

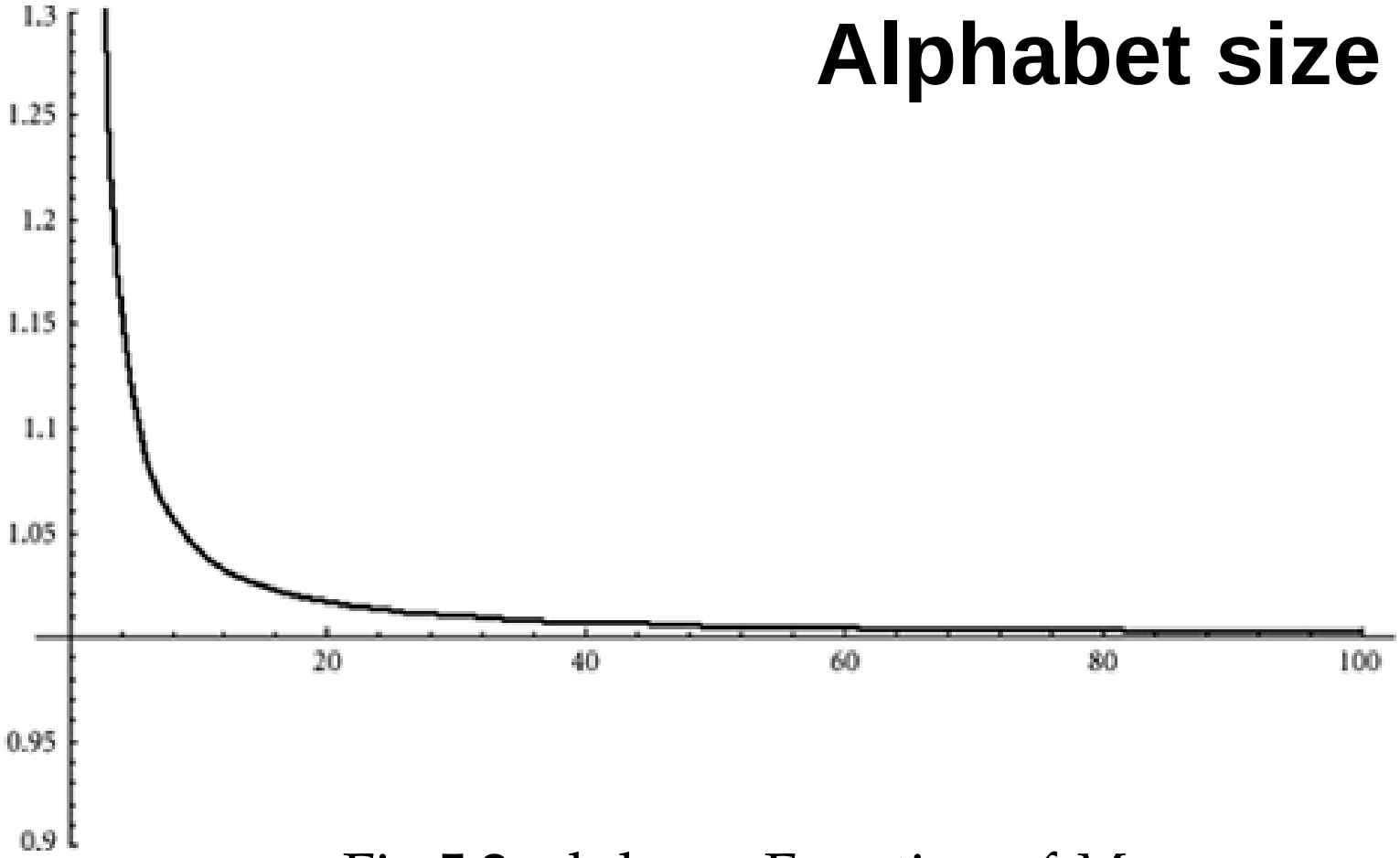
Generalized Zipf

$$p_k = \frac{C}{(\tilde{r} + B)^{\alpha}}$$

- [Mandelbrot,53]

$$p_k = \frac{1}{M}\left(\frac{2(M-1)\tilde{r}}{M+1} + 1\right)^{\frac{-\ln(M+1)}{\ln M}}$$

Alphabet size



- Fig 5.2 alpha as Function of *M*

Connection to generalize Pareto Distribution

- NEW: Post-FOA!
- Beta distribution
- Pareto-Feller
- 'Classical' Pareto

NEW: Post-FOA!

- [G. Amati, Bordoni Foundation, Rome; Keith van Rijsbergen, 2002]
- Based on [B. Arnold, 1983]

Beta distribution

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

$$f_Y(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

Pareto-Feller

$$\Pr(W > w) = \Pr\left(U > \left(\frac{w-\mu}{\sigma}\right)^{1/\gamma}\right)$$

$$f_W(w) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(1+\left(\frac{w-\mu}{\sigma}\right)^{1/\gamma}\right)^{-(\alpha+\beta)} \left(\frac{w-\mu}{\sigma}\right)^{\frac{\beta}{\gamma}-1}$$

- Substitute y=1/(u+1)
- U monotonically decreasing

‘Classical’ Pareto

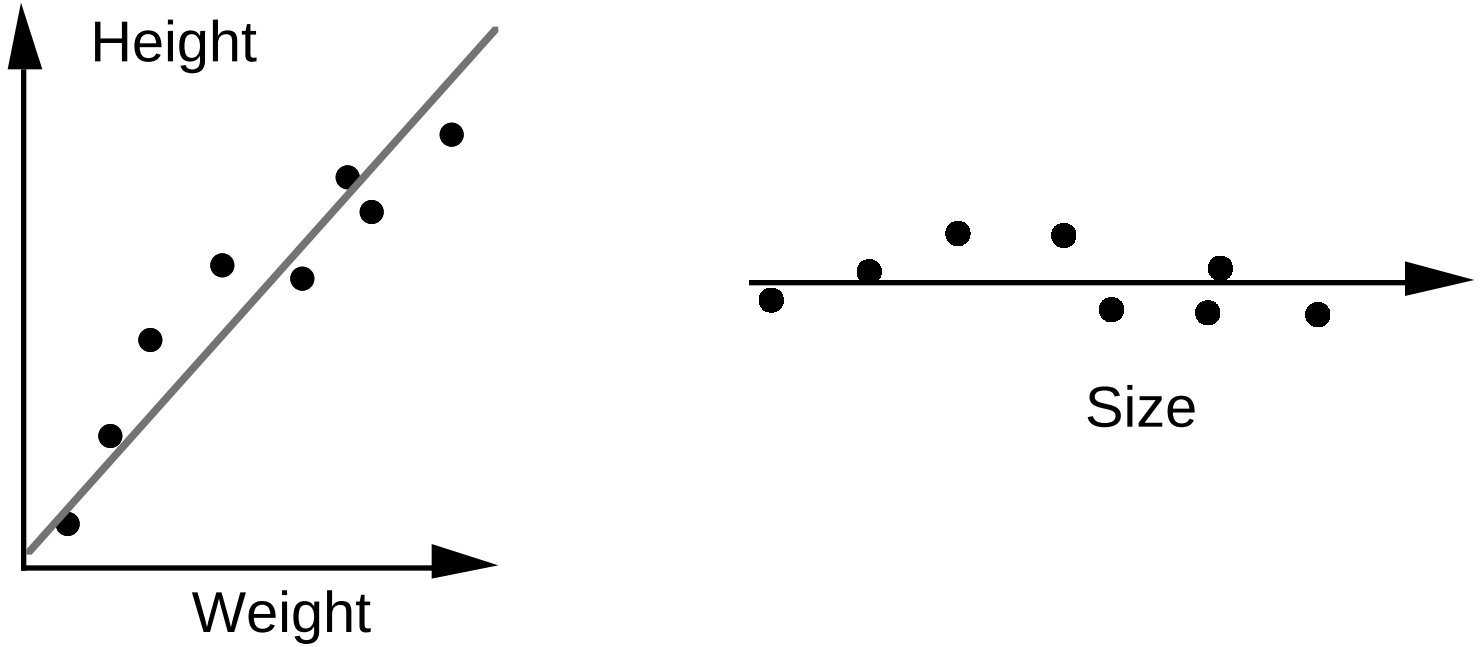
$$W(\mu,\sigma,\gamma,\alpha,\beta)$$

$$\text{Classical}: \mu=\sigma,\gamma=1,\alpha,\beta=1$$

$$\Pr(W>x)=\int_{\sigma}^{x}\frac{\alpha}{\sigma}\left(\frac{w}{\sigma}\right)^{-(\alpha+1)}dw$$

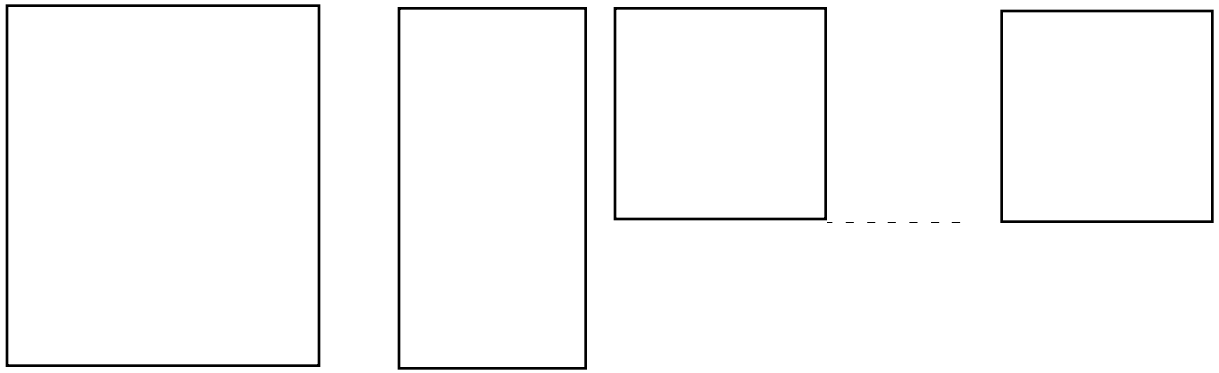
$$=1-\left(\frac{x}{\sigma}\right)^{-\alpha}$$

5.2.1 A Simple Example



- Fig 5.3: Weight and Height Data Reduction

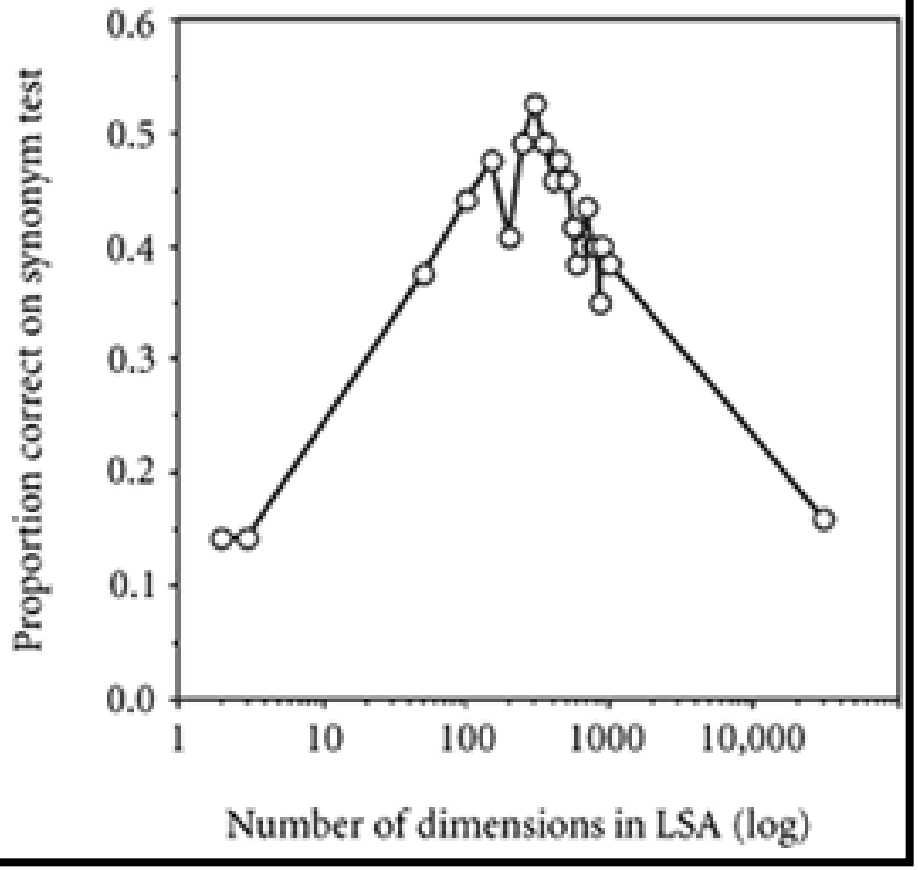
SVD Decomposition



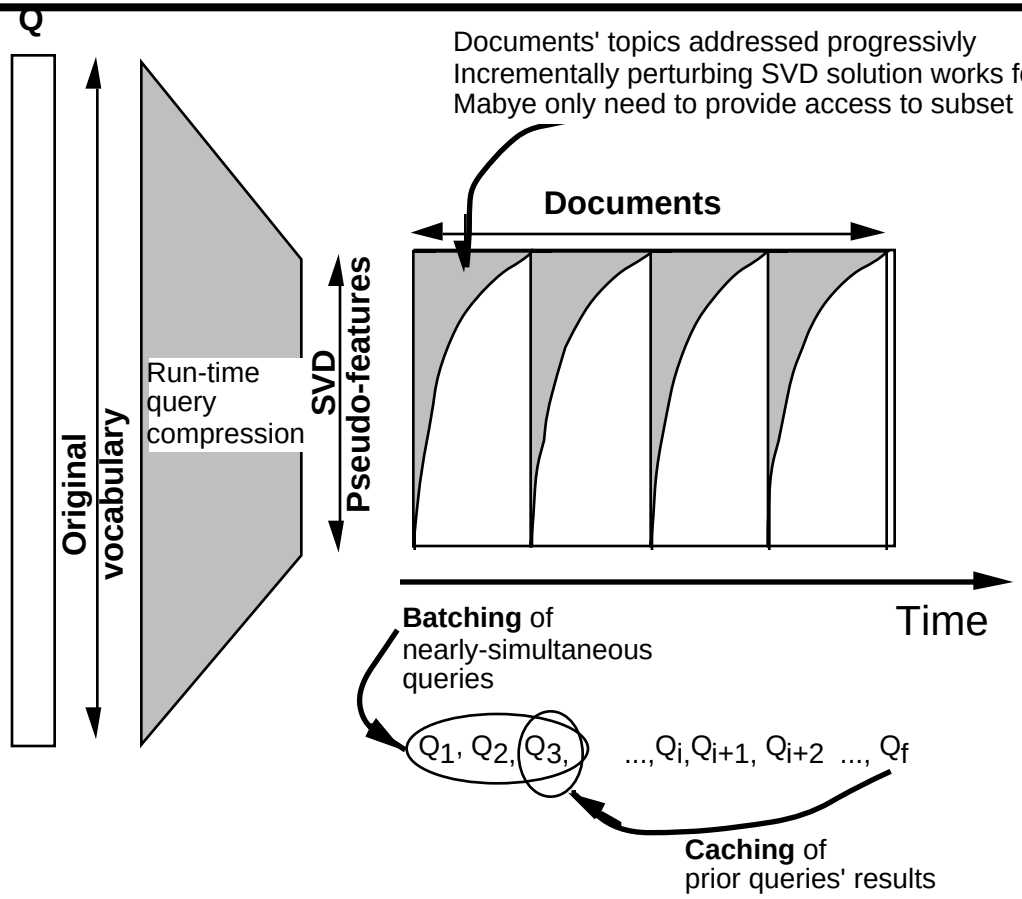
- Reduce dimensionality to 'subspace' which captures primary variation
- Optimal, in terms of minimal (least mean square) error in

5.2.4 How many dimensions?

- Empirically: O(100)
- [Landauer&Dumais'97]



5.2.6 Computational Considerations



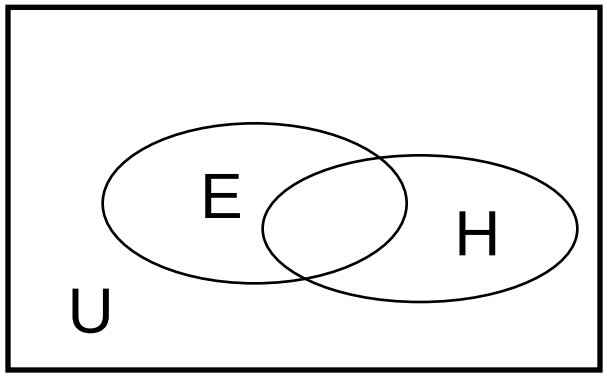
5.5 Probabilisitic Retrieval

- Background
- 5.5.1 Probability Ranking Principle
- 5.5.2 Bayesian Inversion
- 5.5.3 Odds Calculation
- 5.5.4 Binary Independence Model
- 5.5.5 Linear Discriminators
- 5.5.6 Cost Analysis
- 5.5.7 Bayesian networks

Background

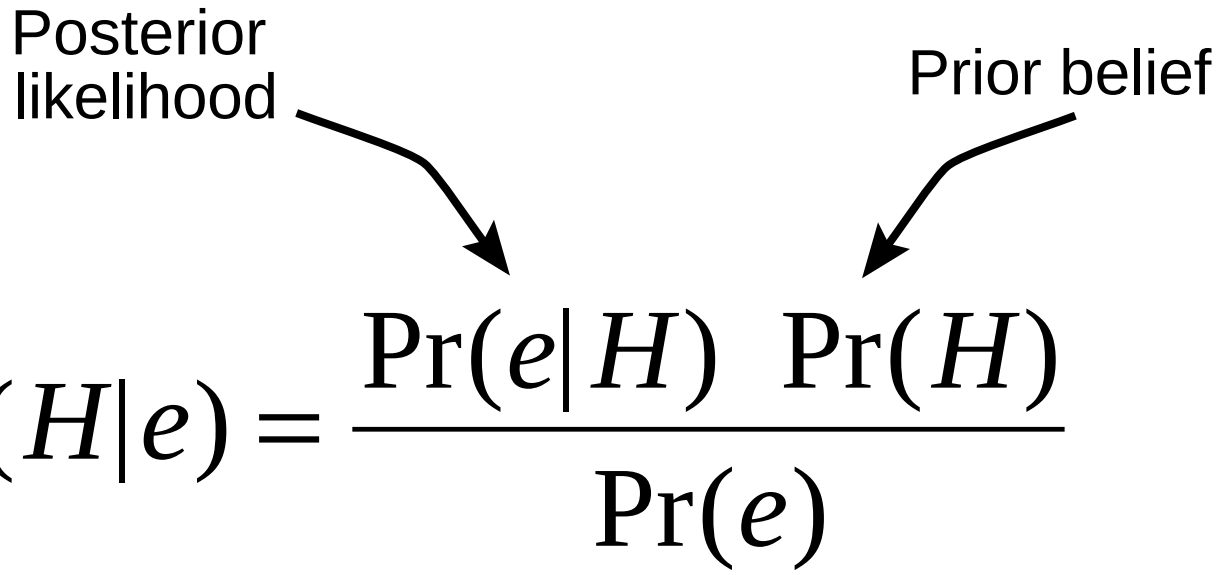
- Conditional probability
- Bayes Rule
- Bayes with partitioning hypotheses
- Problems with Bayes

Conditional probability



$$\Pr(H|E) \equiv \frac{\Pr(H \text{ I } E)}{\Pr(E)}$$

Bayes Rule



Bayes with partitioning hypotheses

$$\Pr(E) = \sum_i \Pr(E|H_i)\ \Pr(H_i)$$

$$\Pr(E \& H_i) = \Pr(H_i|E)\ \Pr(E)$$

$$\Pr(H_j|E) = \frac{\Pr(E|H_j)\Pr(H_j)}{\sum_i \Pr(E|H_i)\ \Pr(H_i)}$$

Problems with Bayes

- Assumes $\{H_i\}$ exhaust U
- Assumes disjoint H_i, H_j
- Otherwise, entire powerset $2^{\{Hi\}}$ must be considered
- Hard (expensive) to obtain prior and posterior estimates
- Changes require global updates, since $\Sigma P_i = 1$

5.5.1 Probability Ranking Principle

- Assuming that the relevance of a document is independent of every other document in the collection [Robertson'77]
- Ranking documents in decreasing order of **probability of relevance** (with respect to any particular query) is optimal
  - But "... that the *probability* of relevance means something be objected to on the same grounds that one might object to probability of Newton's Second Law of Motion being the case.... the probability is either one or zero depending on whether it is true or false." [vanR, p. 127-8]

5.5.2 Bayesian Inversion

- Defining terms
- Bayes Rule, applied to Relevance

Defining terms

$$\Pr(\text{Rel} \mid \underline{x}, \underline{q}) \quad \textit{where}$$

$$\underline{x} = \{\text{features of document}\}$$

$$\underline{q} = \{\text{features of query}\}$$

Bayes Rule, applied to Relevance

$$\Pr(Rel|\underline{x}) = \frac{\Pr(\underline{x}|Rel)\Pr(Rel)}{\Pr(\underline{x})}$$

$$\Pr(\overline{Rel}|\underline{x}) = \frac{\Pr(\underline{x}|\overline{Rel})\Pr(\overline{Rel})}{\Pr(\underline{x})}$$

5.5.3 Odds Calculation

$$Odds(\text{Rel} \mid \underline{x}) = \frac{\Pr(\text{Rel} \mid \underline{x})}{\Pr(\overline{\text{Rel}} \mid \underline{x})}$$

$$= \frac{\Pr(\text{Rel})}{\Pr(\overline{\text{Rel}})} \cdot \frac{\Pr(\underline{x} \mid \text{Rel})}{\Pr(\underline{x} \mid \overline{\text{Rel}})}$$

$$= Odds(\text{Rel}) \cdot \frac{\Pr(\underline{x} \mid \text{Rel})}{\Pr(\underline{x} \mid \overline{\text{Rel}})}$$

Balance evidence of irrelevance, too

Assumptions

- Assume binary features: $\mathbf{x_i = \{0,1\}}$
- Assume document features are independent(!?)

$$\Pr(\underline{\mathbf{x}} \mid \text{Rel}) = \prod_i \Pr(x_i \mid \text{Rel})$$

- More precisely, only need to assume ir/relevant ratios independent

$$\frac{\Pr(\underline{\mathbf{x}} \mid \text{Rel})}{\Pr(\underline{\mathbf{x}} \mid \overline{\text{Rel}})} = \prod_i \frac{\Pr(x_i \mid \text{Rel})}{\Pr(x_i \mid \overline{\text{Rel}})}$$

Bayesian Odds in BIM

$$Odds(\text{Rel} \mid \underline{x}) = Odds(\text{Rel}) \cdot \prod_i \frac{\Pr(x_i \mid \text{Rel})}{\Pr(x_i \mid \overline{\text{Rel}})}$$

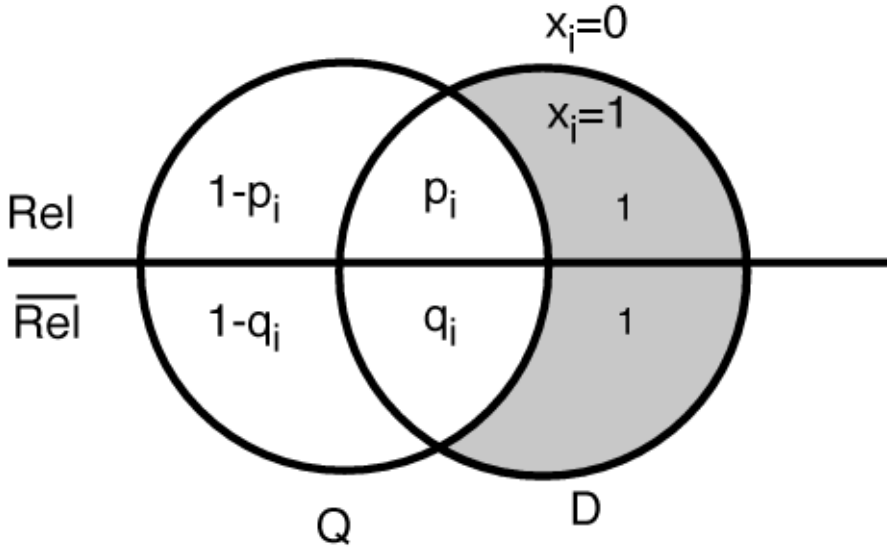
- Useful definitions:

$$p_i \equiv \Pr(x_i = 1 \mid \text{Rel})$$

$$q_i \equiv \Pr(x_i = 1 \mid \overline{\text{Rel}})$$

$$1 - p_i = \Pr(x_i = 0 \mid \text{Rel})$$

Split according to features present in (current) document



$$Odds(\text{Rel} \mid \underline{x}) = Odds(\text{Rel}) \cdot \prod_{x_i=1} \frac{p_i}{q_i} \cdot \prod_{x_i=0} \frac{1-p_i}{1-q_i}$$

5.5.5 Linear Discriminators

- log(•) useful for converting products to more tractible sums
  - And since it's monotonic, rank won't change
- Assume $p_i = q_i$ for all x_i not in query

$$Rank(\underline{x}) = \sum_{x_i \in (q \cap d)} \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

- "Relevance" weight

$$c_i = \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

5.5.6 Cost Analysis

$$C = \log \frac{\Pr(\text{Rel})}{\Pr(\overline{\text{Rel}})} + \sum_{x_i = 0} \log \frac{1 - p_i}{1 - q_i} + \log \frac{Loss_{R\bar{R}}}{Loss_{\bar{R}R}}$$

Prior

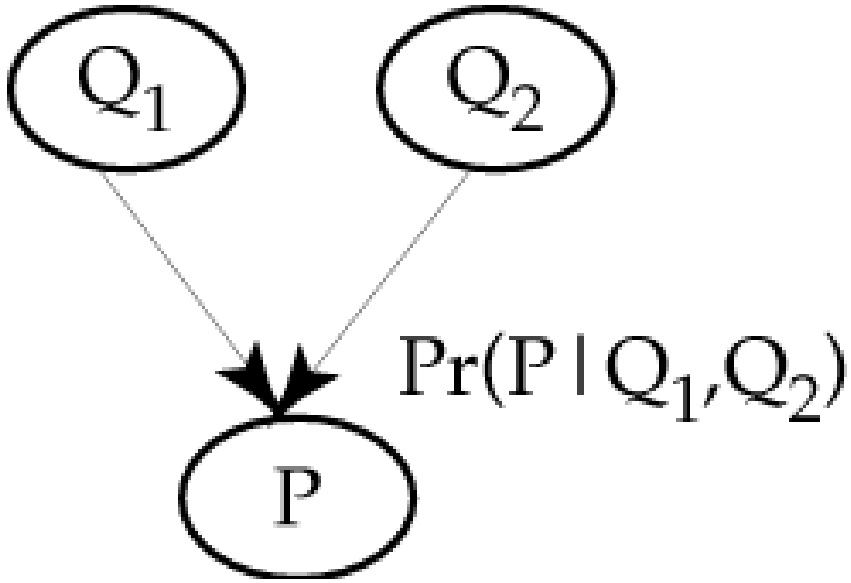
Cost ratio

- A cost model for the Precision/Recall cutoff

5.5.7 Bayesian networks

- Localization of ‘parental’ influences
- Turtle, Croft model
- FOA Network
- Query Modeling
- “Concept Matching” model
- “Concept Matching” (cond.)

Localization of ‘parental’ influences

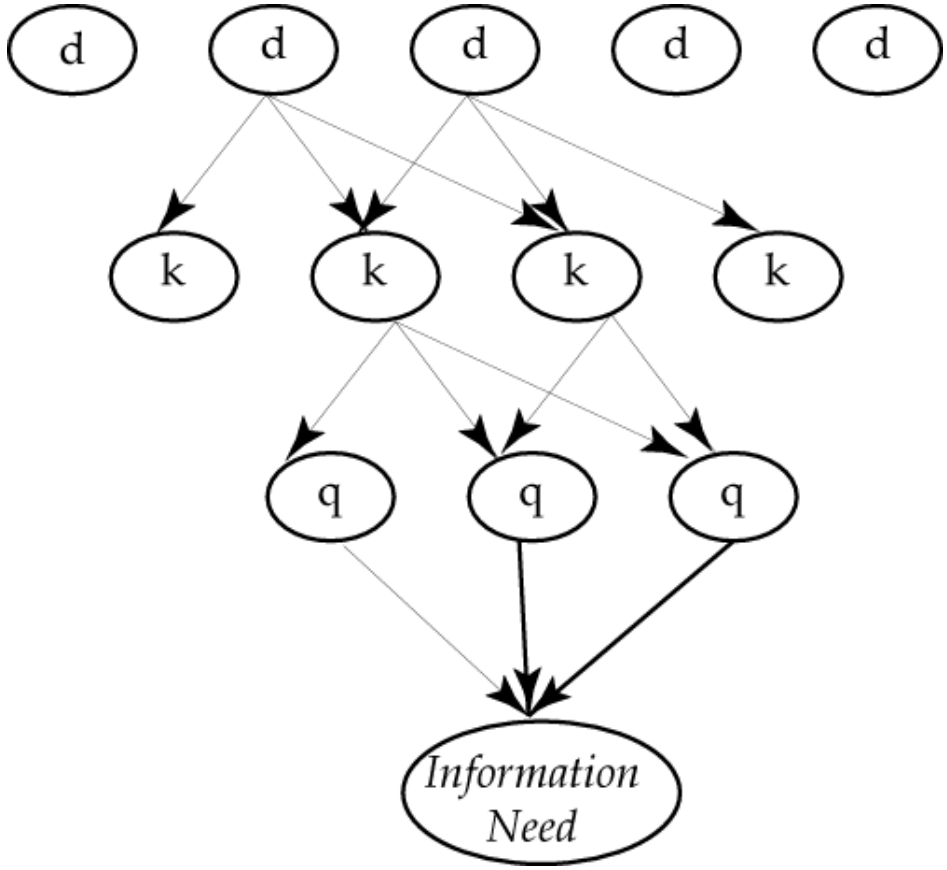


- Fig 5.7

Turtle, Croft model

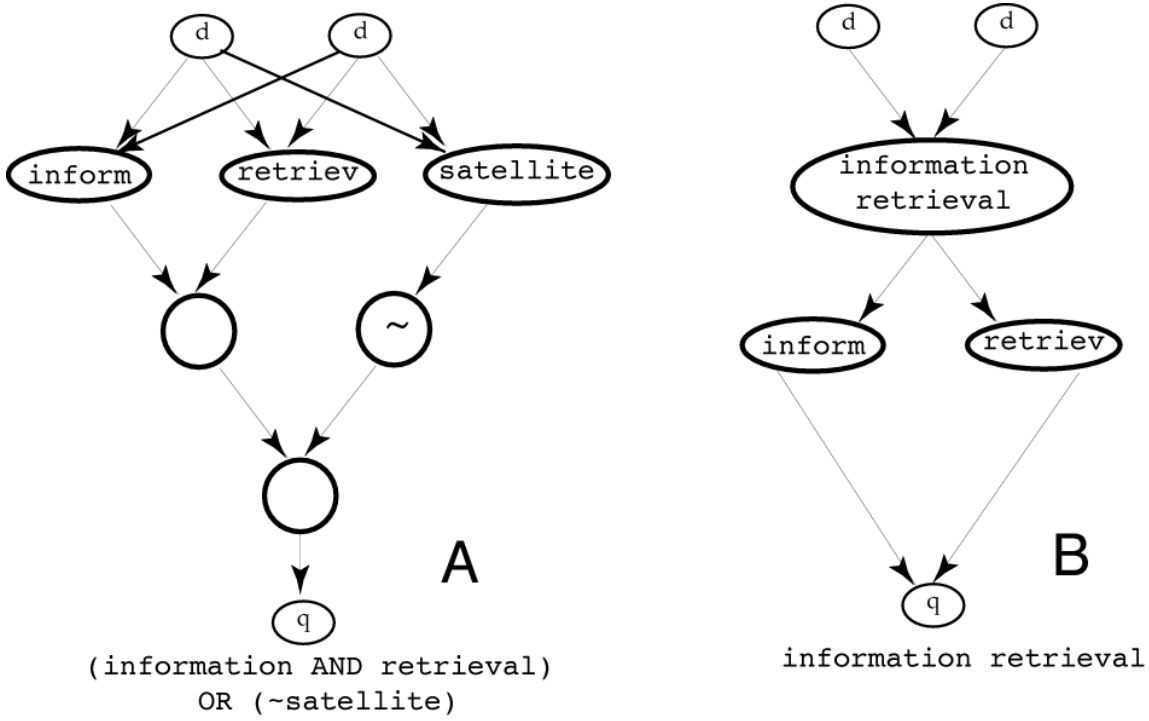
- Doc is observed
- Term is assigned to doc
- Query confirmed
- Doc->KW causal link

FOA Network



- Fig 5.8

Query Modeling



- Fig 5.10 Boolean, phrases

“Concept Matching” model

- ala [Ribiero, Wong]
- Sample space = terms
- Query given as evidence
- Doc is believed to be relevant
- Term is selected/made active

$$P(d|q) = \alpha \sum_u P(d|u)P(q|u)P(u)$$

$$P(q|u) = \begin{cases} 1 \;\; iff \;\; u = q \\ 0 \;\; otherwise \end{cases}$$

$$P(d|u) = \frac{\bar{d} \bullet \bar{u}}{|\bar{d}| \bullet |\bar{u}|}$$

“Concept Matching” (cond.)

- Link from KW to Doc!
- Causal influence reversed!?
- Fig 5.9