

Information Retrieval Systems

Probabilistic Model:

Probability, Information, and Relevance

Weimao Ke

wk@drexel.edu

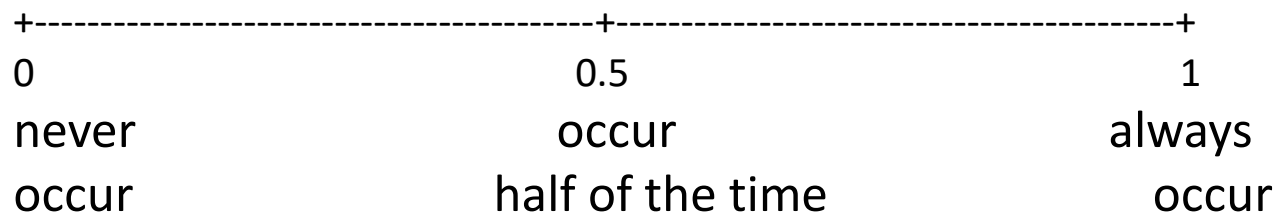
This lecture

- Probability Basics
- Information Theory
- Probabilistic Ranking Principle (PRP)
- Probabilistic Model
 - Focuses on the Binary Independence model

PROBABILITY BASICS

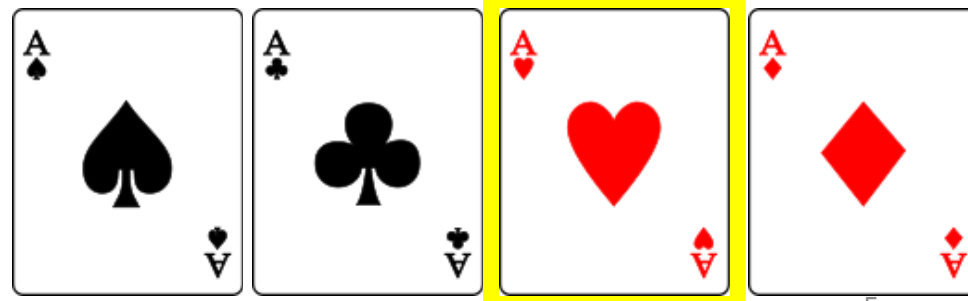
Probability: Definitions & Notations

- Probability is a numerical measure of the likelihood of a given event.
 - number of desirable outcome / total number of possible outcomes
 - $P(A)$ denotes the probability that event A will occur
 - $P(A') \equiv P(A^c) \equiv P(\bar{A}) \equiv$ The probability that A will not occur
 - $P(A') = 1 - P(A)$
- Any probability is always between 0 and 1
 - $0.00 \leq P(A) \leq 1.00$
 - The probability that any event A will occur varies from 0 to 1.



Probability: Definitions & Notations

- $P(A \text{ or } B) \equiv P(A \cup B)$
 - \equiv The probability that either A or B will occur
 - Mutually Exclusive events
 - Events A and B cannot occur simultaneously
 - e.g.
 - A = coin toss turns up a head
 - B = coin toss turns up a tail
 - Not Mutually Exclusive events
 - Events A and B can occur simultaneously
 - e.g.
 - A = a card drawn is an ace
 - B = a card drawn is a heart



Probability: Definitions & Notations

- $P(A, B) \equiv P(A \text{ and } B) \equiv P(A \cap B)$
 - \equiv The probability that both A and B will occur
 - Independent Events
 - The occurrence of A is independent of the occurrence of B
 - e.g.
 - A = the first coin toss turns up a head
 - B = the second coin toss turns up a tail
 - Dependent Events
 - The occurrence of A is not independent of the occurrence of B
 - e.g. , 5 red balls and 5 blue balls in a box
 - A = the first ball drawn is red
 - B = the second ball drawn (without replacement) is red
- $P(B|A) \equiv$ The probability that B will occur given A has occurred.
 - The *conditional probability* of B given A

Probability: Examples

- $P(A \text{ or } B) \equiv P(A \cup B)$
 - \equiv The probability that either A or B will occur
- Mutually Exclusive events
 - Events A and B cannot occur simultaneously
 - $P(A \cup B) = P(A) + P(B)$
 - What is the probability of rolling a die and getting 1 or 6?
 - A = The die rolls a 1
 - B = The die rolls a 6
 - $P(A) = 1/6$
 - $P(B) = 1/6$
 - $P(A \cup B) = 1/6 + 1/6 = 1/3$

Probability: Examples

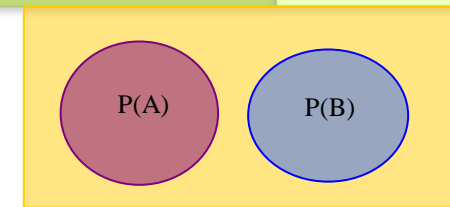
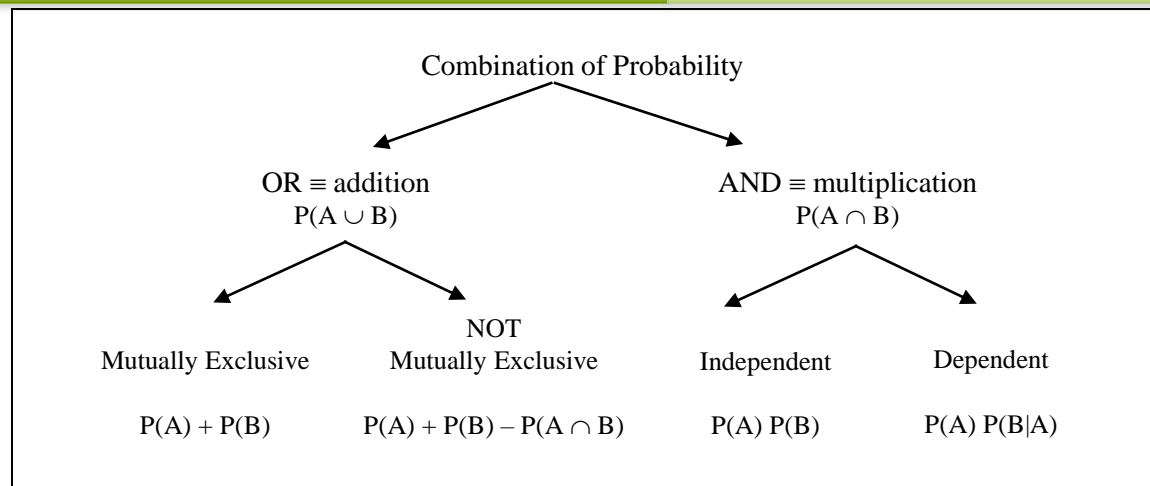
- $P(A \text{ or } B) \equiv P(A \cup B)$
 - \equiv The probability that either A OR B will occur
- Not Mutually Exclusive events
 - Events A and B can occur simultaneously
 - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
 - What is the probability that a card drawn from a deck is an ace or spade?
 - A = The card is an ace
 - B = The card is a spade
 - $P(A) = 4/52$
 - $P(B) = 13/52$
 - $P(A \cap B) = 1/52$
 - $P(A \cup B) = 4/52 + 13/52 - 1/52 = 16/52 = 4/13$

Probability: Examples

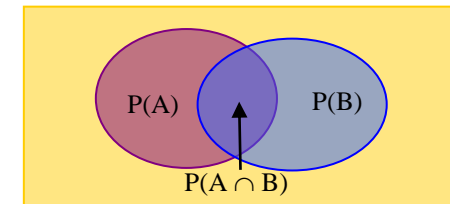
- $P(A, B) \equiv P(A \text{ and } B) \equiv P(A \cap B)$
 - \equiv The probability that both A and B will occur
- Independent Events
 - The occurrence of A is independent of the occurrence of B
 - $P(A \cap B) = P(A) P(B)$
 - What is the probability that fair coin tosses will come up heads twice in a row?
 - A = a head on the first toss
 - B = a head on the second toss
 - $P(A) = P(B) = 1/2$
 - $P(A \cap B) = 1/2 * 1/2 = 1/4$

Probability: Examples

- Dependent Events
 - The occurrence of A is not independent of the occurrence of B
- $P(A \cap B) = P(A) P(B|A) = P(B) P(A|B)$
 - Given 5 pink balls and 5 green balls, what is the probability that two balls drawn are both pink?
 - A = the first ball is red
 - B = the second ball is red
 - $P(A) = 5/10 = 1/2$
 - $P(B|A) = 4/9$
 - $P(A \cap B) = \frac{1}{2} * \frac{4}{9} = \frac{4}{18}$



Mutually Exclusive Events



Not Mutually Exclusive Events

You only need to understand combinations of two probabilities (above).

The following is for your information only (optional):

- Addition (A OR B OR C ...)
- Mutually Exclusive
 - $P(A \cup B \cup C) = P(A) + P(B) + P(C)$
- Not Mutually Exclusive
 - $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$
 - $AB = A \cap B$
 - $P(A \cup B \cup C) = P(AB \cup C)$
 - $= P(AB) + P(C) - P(AB \cap C)$
 - $P(AB) = P(A) + P(B) - P(A \cap B)$
 - $P(AB \cap C) = P((A \cup B) \cap C)$
- Multiplication (A AND B AND C)
- Independent
 - $P(A, B, C) = P(A \cap B \cap C) = P(A) P(B) P(C)$
- Dependent
 - $P(A, B, C) = P(A \cap B \cap C)$
 - $= P(A) P(B|A) P(C|(A \cap B))$
 - $AB = A \cap B$
 - $P(A \cap B \cap C) = P(AB \cap C) = P(AB) P(C|AB)$
 - $P(AB) = P(A) P(B|A)$

Probability: Examples

- $P(A | B)$
 - \equiv The probability that A will occur given B has occurred.
- The *conditional probability* of A given B
 - $P(A | B) = P(A \cap B) / P(B)$
 - Derived from $P(A, B) = P(A \cap B) = P(B) P(A | B)$
 - At a restaurant, 90% of the customers order a burger. If 72% of the customers order a burger and fries, what is the probability that **a customer who orders a burger will also order fries?**
 - A = a customer orders fries
 - B = a customer orders a burger
 - $P(B) = 0.9$, $P(A \cap B) = 0.72$
 - $P(A | B) = P(A \cap B) / P(B) = 0.72 / (90/100) = 72/90 = 8/10 = 80\%$

Probability: Bayesian Theorem

- Bayesian Theorem

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

- $P(A)$ is the probability A.
 - $P(A | B)$ is the conditional probability of A, given B.
 - $P(B | A)$ is the conditional probability of B, given A.
 - $P(B)$ is the probability of B.
-
- Important to probabilistic IR (later in this lecture)

Probability: Odds

- Odds is the relative likelihood that an event will happen
 - = probability that the event will happen divided by the probability that it won't

$$O(A) = \frac{P(A)}{1 - P(A)}$$

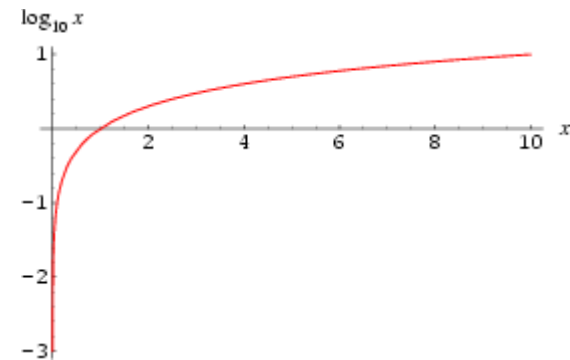
- where $P(A)$ is probability of event A

Information, probability, and uncertainty/entropy

INFORMATION THEORY

Some Basics on Logarithm

- $\log_b x = y$
 - Logarithm of x to base b
 - $x = b^y$
- Common Logarithm
 - Logarithm to base 10
 - $\log_{10} x = \log x$
- Natural Logarithm
 - Logarithm to base e , where $e = 2.718281828\dots$
 - $\log_e x = \ln x$
- Properties
 - $\log x_1 x_2 = \log x_1 + \log x_2$
 - $\log \frac{x_1}{x_2} = \log x_1 - \log x_2$
 - $\log x^k = k \log x$
- Log Transformation
 - $\log_b x = \frac{\log_c x}{\log_c b}$
 - $x = b^y, \log_b x = y$
 - $\log_c x = \log_c b^y = y \log_c b = \log_b x \log_c b$
 - $\log_2 x = \frac{\log x}{\log 2}$
 - $x = 2^y, \log_2 x = y$
 - $\log_{10} x = \log_{10} 2^y = y \log_{10} 2 = \log_2 x \log_{10} 2$



Logarithm Examples

$$\log_{10} 1 = 0$$

$$\log_{10} 10 = \log_{10} 10^1 = 1$$

$$\log_{10} 100 = \log_{10} 10^2 = 2$$

$$\log_{10} 1000 = \log_{10} 10^3 = 3$$

$$\log_{10} 1000 = \log_{10} 10^3 = 3 * \log_{10} 10 = 3 * 1 = 3$$

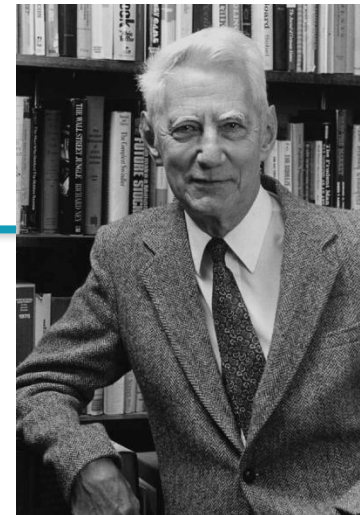
$$\log_{10} (10 * 100) = \log_{10} 10 + \log_{10} 100 = 1 + 2 = 3$$

$$\log_{10} (1000 / 10) = \log_{10} 1000 - \log_{10} 10 = 3 - 1 = 2$$

Understanding these examples should provide a good foundation on logarithm for you to complete this week's assignment.

Information Theory

- Claude Shannon (1948)
 - Title: *A Mathematical Theory of Communication*
 - Connects information with uncertainty (entropy)
- Important assumption (which you may not agree:)
 - Receiving information (I) reduces uncertainty/entropy (H).
 - Or, information is the decrease of uncertainty.
- Suppose
 - Your initial uncertainty about something is H_0
 - Your uncertainty after getting information is H_1
 - $H_1 = 0$ if you are certain now (it is known and there is no uncertainty)
- Then, the amount of information you've received is
 - $I = H_0 - H_1$
 - $= H_0$ if you are completely certain now ($H_1=0$)



Claude Shannon

Information Theory – entropy

- Now how do we measure uncertainty/entropy
 - then we' ll be able to measure information
- According to Shannon (1948), entropy is:

$$H = - \sum_{i=1}^M P_i \log P_i$$

- given M choices (different potential outcomes)
- P_i is the probability of the i^{th} choice (the likelihood that the i^{th} will come true)

Information Theory – Example

- An example about election:
 - Three candidates 1, 2, and 3 running for office
 - The probability of each candidate being elected is:
 - $P_1 = 0.1$, $P_2 = 0.5$, $P_3 = 0.4$. Apparently, they should add up to 1.0.
 - **Before** the result is out, the uncertainty about the election is:

$$\begin{aligned}
 H_0 &= -\sum_{i=1}^3 P_i \log P_i = -(P_1 \log P_1 + P_2 \log P_2 + P_3 \log P_3) \\
 &= -(0.1 \log 0.1 + 0.5 \log 0.5 + 0.4 \log 0.4) = -(-0.41) = 0.41
 \end{aligned}$$

- **After** election, result/information is known:
 - No matter which candidate wins, there is no uncertainty

$$H_1 = 0$$

- The amount of information about election is:
 - $I = H_0 - H_1 = H_0 = 0.41$

Information Theory – Example

- Election example continued
 - Case 1: suppose it is a tight election, three candidates have the equal chances to win: $P_1 = P_2 = P_3 = 1/3$
 - Intuitively, election result is harder to predict
 - Uncertainty should be larger
 - Yes, the entropy formula tells the same story:

$$H_0 = 0.48 > 0.41$$

- Case 2: suppose it is an easy election for candidate 2, who has 90% likelihood to win
 - Say $P_1=0.01$, $P_2=0.9$, $P_3=0.09$
 - Uncertainty should be smaller
 - Yes, the number is consistent: $H_0 = 0.15 \ll 0.41$

Information Theory - importance

- Shannon's equation was a very important discovery:

$$H = -\sum_{i=1}^M P_i \log P_i$$

- In a sense, its significance to communication/computing/information sciences can be likened to that of Einstein's $E=MC^2$ to physics.
 - It enables us to measure the amount of information and estimate capacities needed to communicate and/or store that amount of information.
 - It provides guidance on information compression.
 - It lays foundations for digital information technologies (dealing with *bits*)
 - ...

Information Theory - characteristics

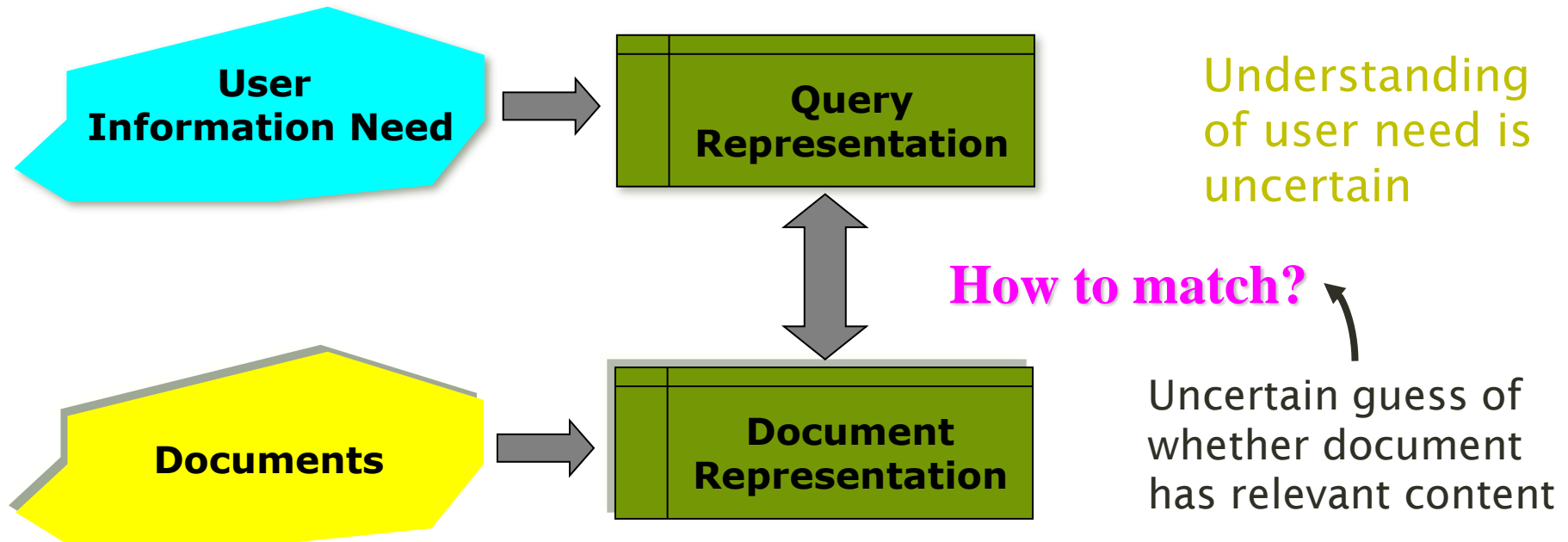
- Why is this *THE* formula?

$$H = - \sum_{i=1}^M P_i \log P_i$$

- What makes it meaningful?
- Three important characteristics, among others:
 - H is continuous in the p_i . That is, smooth changes in probability values cause smooth (not sudden/dramatic) changes of H.
 - H is larger given a larger M – that is, in general, there is more uncertainty given more choices (some qualifications).
 - If a choice is broken down into two successive choices, the original H is the weighted sum of the individual values of H.

PROBABILITY RANKING PRINCIPLE

Why probabilities in IR?



In traditional IR systems, matching between each document and query is attempted in a semantically imprecise space of index terms.

Probabilities provide a principle foundation for *uncertainty* reasoning.
Can we use probabilities to quantify our uncertainties?

Probabilistic IR topics

- *Probabilistic methods are one of the oldest but also one of the currently hottest topics in IR.*

The document ranking problem

- We have a collection of documents
- User issues a query
- A list of documents needs to be returned
- **Ranking method is core of an IR system:**
 - In what order do we present documents to the user?
 - The “best” document to be first, second best second,
- **Idea: Rank by probability of relevance of the document w.r.t. information need (query)**
 - $P(\text{relevant} | \text{document}_i, \text{query})$
 - Another notation: $P(R | d, q)$
 - Yet another notation: $P(R | X, q)$, or simply $P(R | X)$
 - Where d is represented by vector of terms: $X = (x_1, x_2, \dots, x_k)$

The Probability Ranking Principle

“If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of **decreasing probability of relevance** to the user who submitted the request, where the probabilities are estimated **as accurately as possible on the basis of whatever data** have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data.”



Karen Spärck Jones



Stephen Robertson

Focuses on the Binary Independence Model

PROBABILISTIC MODEL

Probabilistic Ranking

Basic concept:

"For a given query, if we know some documents that are relevant, terms that occur in those documents should be given greater weighting in searching for other relevant documents.

By making assumptions about the distribution of terms and applying Bayes Theorem, it is possible to derive weights theoretically."



van Rijsbergen

Probability Ranking Principle

Let x be a document in the collection.

Let R represent **relevance** of a document w.r.t. given (fixed) query q and let NR represent **non-relevance**.

Need to find $p(R/x)$ - probability that a document x is **relevant**.

more formally, written as $p(R|x,q)$

Probability of Being Relevant

- $p(R|x)$ – probability that document x is relevant
- $p(NR|x)$ – probability that document x is nonrelevant

- $p(R|x) + p(NR|x) = 1$

$$p(R|x) = \frac{p(x|R)p(R)}{p(x)}$$

$$p(NR|x) = \frac{p(x|NR)p(NR)}{p(x)}$$

Bayesian Theorem

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

- $p(x/R)$ – probability of a retrieved relevant document being x .
- $p(x/NR)$ - probability of a retrieved non-relevant document being x .

Probability Ranking Principle

- How do we compute all those probabilities?
 - Do not know exact probabilities, have to use estimates
 - Binary Independence Retrieval (BIR) – which we discuss later today – is the simplest model
- Assumptions (questionable though)
 - “Relevance” of each document is independent of relevance of other documents.
 - Really, it’s bad to keep on returning **duplicates**
 - Boolean model of relevance
 - That one has a single step information need
 - Seeing a range of results might let user refine query
- **But for a simplified discussion, we only use the Binary Independence model for probabilistic retrieval.**

Probabilistic Retrieval Strategy

- Estimate how terms contribute to relevance
 - How do things like term frequency (tf), document frequency (df), and document length influence your judgments about document relevance?
- Combine to find document relevance probability
- Order documents by decreasing probability

Binary Independence Model

- Traditionally used in conjunction with PRP
- “**Binary**” = **Boolean**: documents are represented as binary incidence vectors of terms:
 - document : $\vec{x} = (x_1, \dots, x_n)$
 - $x_i = 1$ if term i is present in document x .
- “**Independence**”: terms occur in documents independently
- Different documents can be modeled as same vector

Binary Independence Model

- **“Binary” = Boolean**: documents are represented as binary incidence vectors of terms:
 - document : $\vec{x} = (x_1, \dots, x_n)$
 - $x_i = 1$ if term i is present in document x .
- **“Independence”**: terms occur in documents independently



“Head”



“Five”

... ..

A document: $X = (\quad 1 \quad \quad 1 \quad \quad 0 \quad 0 \quad \dots)$

$$P(X) = P(\text{“Head”}) \times P(\text{“Five”}) \times \dots = \prod p(x_i)$$

Binary Independence Model

- Queries: binary term incidence vectors
- Given query q ,
 - for each document d need to compute $p(R|q,d)$.
 - replace with computing $p(R|q,x)$ where x is binary term incidence vector representing d
 - Interested only in ranking
- Will use odds and Bayes' Rule:

$$O(R|q,\vec{x}) = \frac{p(R|q,\vec{x})}{p(NR|q,\vec{x})} = \frac{\frac{p(R|q)p(\vec{x}|R,q)}{p(\vec{x}|q)}}{\frac{p(NR|q)p(\vec{x}|NR,q)}{p(\vec{x}|q)}}$$

Recall the odds equation?

Bayesian Theorem

Binary Independence Model

you may skip equations here

$$O(R | q, \vec{x}) = \frac{p(R | q, \vec{x})}{p(NR | q, \vec{x})} = \underbrace{\frac{p(R | q)}{p(NR | q)}}_{\text{Constant for a given query}} \times \underbrace{\frac{p(\vec{x} | R, q)}{p(\vec{x} | NR, q)}}_{\text{Needs estimation}}$$

- Using **Independence** Assumption:

$$\frac{p(\vec{x} | R, q)}{p(\vec{x} | NR, q)} = \prod_{i=1}^n \frac{p(x_i | R, q)}{p(x_i | NR, q)}$$

- So : $O(R | q, d) = O(R | q) \times \prod_{i=1}^n \frac{p(x_i | R, q)}{p(x_i | NR, q)}$

Binary Independence Model

you may skip equations here

$$O(R | q, d) = O(R | q) \cdot \prod_{i=1}^n \frac{p(x_i | R, q)}{p(x_i | NR, q)}$$

- Since x_i is either 0 or 1:

$$O(R | q, d) = O(R | q) \times \prod_{x_i=1} \frac{p(x_i = 1 | R, q)}{p(x_i = 1 | NR, q)} \times \prod_{x_i=0} \frac{p(x_i = 0 | R, q)}{p(x_i = 0 | NR, q)}$$

- Let $p_i = p(x_i = 1 | R, q)$; $r_i = p(x_i = 1 | NR, q)$;

- Assume, for all terms not occurring in the query ($q_i=0$) $p_i = r_i$

Then...

This can be changed (e.g., in relevance feedback)

Some more steps...

$$O(R | q, d) = O(R | q) \times \prod_{x_i=1} \frac{p(x_i = 1 | R, q)}{p(x_i = 1 | NR, q)} \times \prod_{x_i=0} \frac{p(x_i = 0 | R, q)}{p(x_i = 0 | NR, q)}$$

Binary Independence Model

you may skip equations here

$$O(R | q, \vec{x}) = \underbrace{O(R | q)}_{\text{All matching terms}} \times \underbrace{\prod_{x_i = q_i = 1} \frac{p_i}{r_i}}_{\text{Non-matching query terms}} \times \underbrace{\prod_{x_i = 0, q_i = 1} \frac{1 - p_i}{1 - r_i}}_{\text{Non-matching query terms}}$$

All matching terms

Non-matching
query terms

Binary Independence Model

you may skip equations here

$$O(R | q, \vec{x}) = \underbrace{O(R | q)}_{\text{All matching terms}} \times \underbrace{\prod_{x_i=q_i=1} \frac{p_i}{r_i}}_{\text{Non-matching query terms}} \times \underbrace{\prod_{\substack{x_i=0 \\ q_i=1}} \frac{1-p_i}{1-r_i}}_{\text{Non-matching query terms}}$$

$$= \underbrace{O(R | q)}_{\text{All matching terms}} \times \underbrace{\prod_{x_i=q_i=1} \frac{p_i(1-r_i)}{r_i(1-p_i)}}_{\text{All query terms}} \times \underbrace{\prod_{q_i=1} \frac{1-p_i}{1-r_i}}_{\text{All query terms}}$$

Binary Independence Model

you may skip equations here

$$O(R | q, \vec{x}) = \boxed{O(R | q)} \times \boxed{\tilde{O}_{x_i=q_i=1} \frac{p_i(1-r_i)}{r_i(1-p_i)}} \times \boxed{\tilde{O}_{q_i=1} \frac{1-p_i}{1-r_i}}$$

Constant for each query

Only quantity to be estimated for rankings

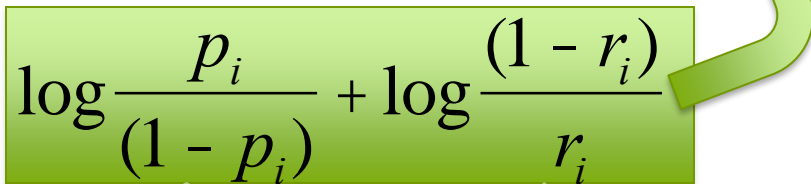
- Retrieval Status Value:

$$RSV = \log \tilde{O}_{x_i=q_i=1} \frac{p_i(1-r_i)}{r_i(1-p_i)} = \sum_{x_i=q_i=1} \log \frac{p_i(1-r_i)}{r_i(1-p_i)}$$

Binary Independence Model

- All boils down to computing Retrieval Status Value:

$$RSV = \sum_{x_i=q_i=1}^K c_i = c_1 + c_2 + \dots + c_k$$

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


$p_i = p(x_i = 1 | R, q)$ probability term i appears in a document relevant to the query

$r_i = p(x_i = 1 | NR, q)$ probability term i appears in a document not relevant to the query

So, how do we compute p_i, r_i values from our data ?

Binary Independence Model

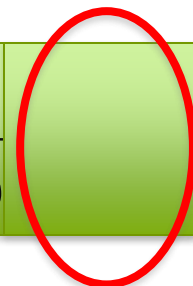
- Estimating RSV coefficients.
- For each term i look at this table of document counts:

Documens	Relevant	Non-Relevant	Total
$X_i=1$	s	$n-s$	n
$X_i=0$	$S-s$	$N-n-S+s$	$N-n$
Total	S	$N-S$	N



• Estimates: $p_i \gg \frac{s}{S}$ $r_i \gg \frac{(n-s)}{(N-S)}$

= 0.5 if we assume
equal chances



Binary Independence Model

$$RSV = \sum_{x_i=q_i=1}^K c_i = c_1 + c_2 + \dots + c_k$$

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

$$= \log \frac{N-n_i}{n_i}$$

$$w_i^{IDF} \sim \log \frac{N-n_i+0.5}{n_i+0.5}$$

Estimation – key challenge

- Many more non-relevant documents than relevant docs
 - If non-relevant documents are approximated by the whole collection, then
 - r_i (prob. of occurrence in non-relevant documents for query) is n/N and
 - $r_i = \log(1 - r_i)/r_i = \log(N - n)/n \approx \log N/n = \text{IDF!}$
- p_i (probability of occurrence in relevant documents) can be estimated in various ways:
 - from relevant documents if know some
 - Relevance weighting can be used in feedback loop
 - Constant, e.g., 0.5 (Croft and Harper combination match)
 - proportional to prob. of occurrence in collection

PRP and BIR

- Getting reasonable approximations of probabilities is possible.
- Requires restrictive assumptions:
 - *term independence*
 - *terms not in query don't affect the outcome*
 - *boolean representation of documents/queries/relevance*
 - *document relevance values are independent*
- Some of these assumptions can be removed
- Problem: either require partial relevance information or only can derive somewhat inferior term weights

IR Theory vs. Practice

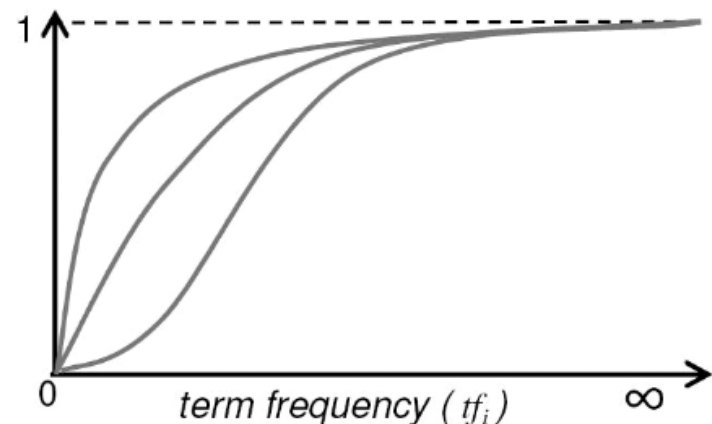
- Standard Vector Space Model:
 - Empirical for the most part; success measured by results
 - Few properties provable
- Binary Independence probabilistic model disadvantages:
 - Has never worked convincingly better in practice
- Probabilistic Model Advantages:
 - Based on a firm theoretical foundation
 - Theoretically justified optimal ranking scheme
- More advanced probabilistic models and ranking methods
 - Okapi BM25, language models, etc.
 - Very effective

Okapi BM25

- Comparable to $TF \cdot IDF$
- BIM: w_i^{IDF} is the IDF component
- What about TF (Term Frequency)?
 - First, saturation to normalize TF to $[0, 1]$

$$w_{di}^{TF} = \frac{tf_{di}}{tf_{di} + k}$$

- Term **i** in document **d**
- Pivot value: k



Okapi BM25

- What about TF (Term Frequency)?
 - Second, document length normalization

$$tf'_{di} \sim \frac{tf_{di}}{(1-b) + b \frac{l_d}{avl}}$$

- l_d is the length of document d
- avl is the average document length in the collection
- b is parameter for the normalization
 - $b = 0$ to switch normalization off
 - $b = 1$ for full document-length normalization

Okapi BM25

- Put together:
- TF: $w_{di}^{TF} = \frac{tf'_{di}}{tf'_{di} + k} = \frac{tf_{di}}{tf_{di} + k \frac{tf_{di}}{(1-b) + b \frac{l_d}{avl}}}$
- IDF: $w_i^{IDF} = \log \frac{N - n_i + 0.5}{n_i + 0.5}$
- BM25: $w_i^{BM25} = w_{di}^{TF} w_i^{IDF}$
 - k and b can be tuned through experiments
 - Good values in general:
 - $0.5 < b < 0.8$
 - $1.2 < k < 2$

Okapi BM25

- Decades of IR research:
 - Firm theoretical foundation
 - Superior experimental results
- Default scoring function in Elasticsearch
- Can be fine tuned and adapted
 - E.g. BM25F

References

MRS Textbook Chapter 11 Probabilistic Information Retrieval.

Claude Shannon (1948). A Mathematical Theory of Communication.
<http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf>

Schneider TD (2010): information Theory Primer (read either the link below or the attached PDF file):<http://alum.mit.edu/www/toms/paper/primer/>

S. E. Robertson and K. Spärck Jones. 1976. Relevance Weighting of Search Terms. *Journal of the American Society for Information Sciences* 27(3): 129–146.

C. J. van Rijsbergen. 1979. *Information Retrieval*. 2nd ed. London: Butterworths, chapter 6. [Most details of math]
<http://www.dcs.gla.ac.uk/Keith/Preface.html>

References

- H.R. Turtle and W.B. Croft. 1990. Inference Networks for Document Retrieval. *Proc. ACM SIGIR*: 1-24.
- E. Charniak. Bayesian nets without tears. *AI Magazine* 12(4): 50-63 (1991).
<http://www.aaai.org/Library/Magazine/Vol12/12-04/vol12-04.html>
- D. Heckerman. 1995. A Tutorial on Learning with Bayesian Networks. Microsoft Technical Report MSR-TR-95-06
<http://www.research.microsoft.com/~heckerman/>
- N. Fuhr. 2000. Probabilistic Datalog: Implementing Logical Information Retrieval for Advanced Applications. *Journal of the American Society for Information Science* 51(2): 95–110.
- R. K. Belew. 2001. *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW*. Cambridge UP 2001.