Princess Sumaya University for Technology جامعــــة الأميــرة سميّــة للتكنولوجيا

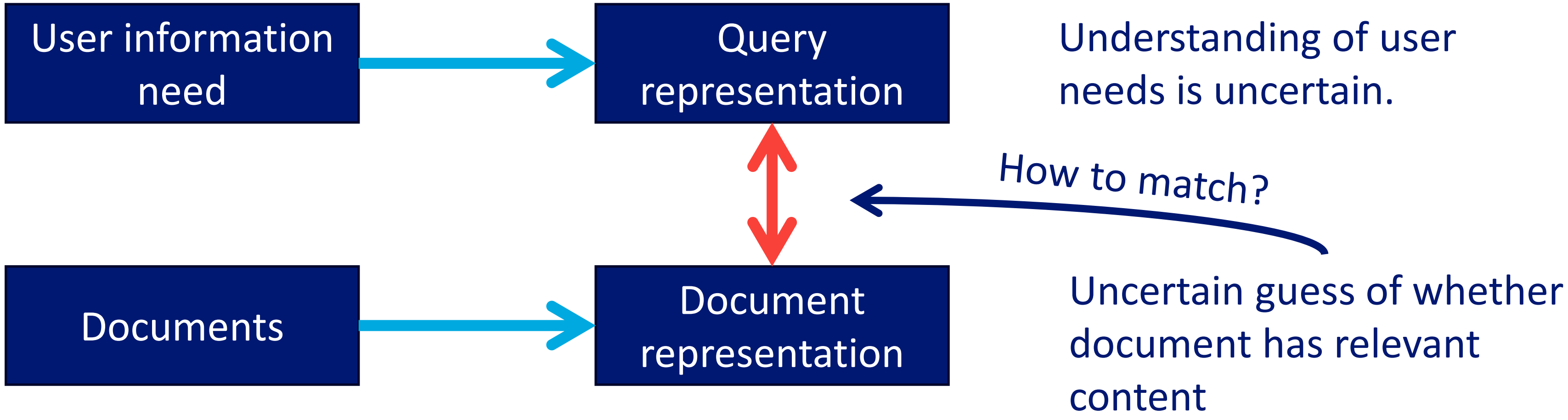
Excellence in Education, since 1991 التميّــز في التعليــم منـذ عام 1991

Probabilistic Information Retrieval

Dr. Bushra Alhijawi

Why Probabilities in IR

- Given a user information need (represented as a query) and a collection of documents, an IR system must determine how well the documents satisfy the query.
- In traditional IR systems, matching between each document and query is attempted in a semantically imprecise space of index terms.



User information need

Query representation

Understanding of user needs is uncertain.

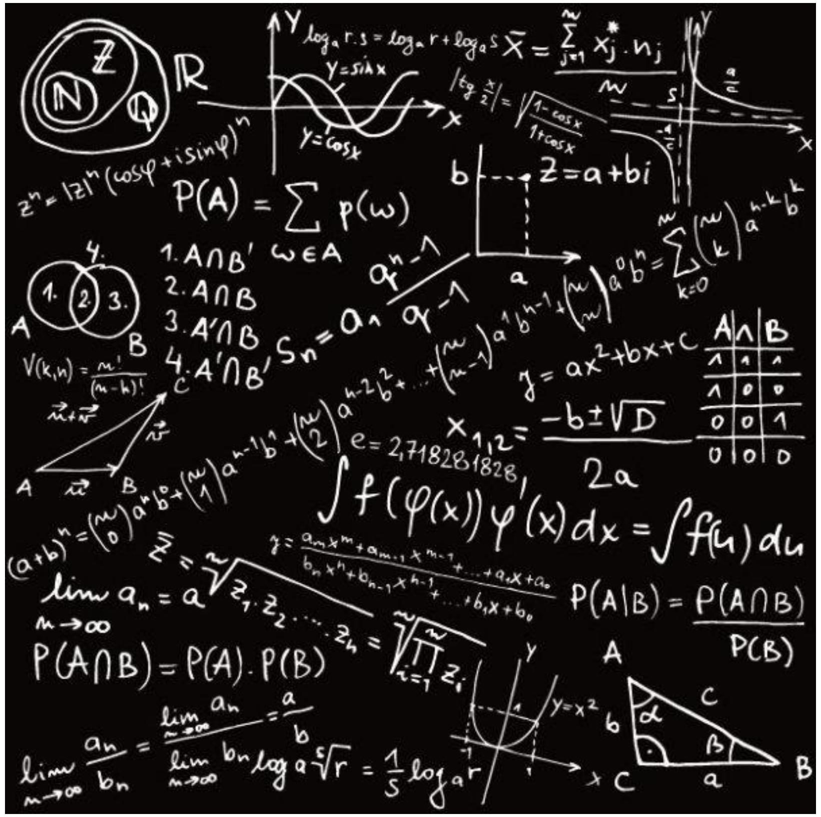
How to match?

Documents

Document representation

Uncertain guess of whether document has relevant content

Why Probabilities in IR



- An IR system has an **UNCERTAIN** understanding of the user query and makes an uncertain guess of whether a document satisfies the query.

- Probability theory provides a principled foundation for such reasoning under uncertainty.

- Probabilistic models exploit this foundation to estimate how likely it is that a document is relevant to a query.

Why Probabilities in IR – Basic Idea

Query	Document	Relevant?
Q1	D1	1
Q1	D2	0
Q1	D3	1
….	….	….
Q1	D1	0
Q1	D2	0
Q2	D1	1
Q2	D2	0
Q3	D1	1
Q4	D2	0

What is the probability that D1 is relevant given that it is retrieved for Q1?

$$Score(d,q) = P(R=1|d,q) = \frac{count(q,d,R=1)}{count(q,d)}$$

$$Score(D1,Q1) = ?$$

$$Score(D2,Q1) = ?$$

$$Score(D3,Q1) = ?$$

Why Probabilities in IR – Basic Idea

Query	Document	Relevant?
Q1	D1	1
Q1	D2	0
Q1	D3	1
….	….	….
Q1	D1	0
Q1	D2	0
Q2	D1	1
Q2	D2	0
Q3	D1	1
Q4	D2	0

What is the probability that D1 is relevant given that it is retrieved for Q1?

$$Score(d,q) = P(R=1|d,q) = \frac{count(q,d,R=1)}{count(q,d)}$$

$$Score(D1,Q1) = \frac{1}{2}$$

$$Score(D2,Q1) = \frac{0}{2}$$

$$Score(D3,Q1) = \frac{1}{1}$$

© Bushra Alhijawi

Why Probabilities in IR – Basic Idea

Query	Document	Relevant?
Q1	D1	1
Q1	D2	0
Q1	D3	1
….	….	….
Q1	D1	0
Q1	D2	0
Q2	D1	1
Q2	D2	0
Q3	D1	1
Q4	D2	0

- What about unseen documents ?!!!
- What about unseen queries ?!!!

How likely the user enter q

$$Score(d,q) = P(R=1|d,q) \approx \underbrace{P(q|\overbrace{d,R=1}^{\text{User likes d}})}_{\text{Approximation}}$$

Assumption

A user formulates a query based on an imaginary relevant document.

Why Probabilities in IR – Basic Idea

Q: news about generative AI

Which of these documents is most likely the imaginary relevant document in the user's mind when the user formulates this query

?

D1: …. **news about** …. $P(Q|D1)$

D2: …. **news about** machine learning and **AI** …. $P(Q|D2)$

D3: …. **news** of **generative AI** …. $P(Q|D3)$

D4: …. **news** of **generative AI** ….. **AI** algorithms …. $P(Q|D4)$

D5: …. **news** of machine learning and **AI** ….. generative …. ….. generative ….. generative …. $P(Q|D4)$

Probability Theory – Review

- Probability theory is the mathematical framework that allows us to analyze chance events in a logically sound manner.
- Basic terminologies associated with probability theory:
  - Probability is the chance of happening or occurrences of an event.
    - This number is always between 0 and 1, where 0 indicates impossibility and 1 indicates certainty.
  - Event → a set of outcomes of an experiment that forms a subset of the sample space.
    - Independent events → Events that are not affected by other events are independent events.
    - Dependent events → Events that are affected by other events are known as dependent events.
  - Random variable → a variable that assumes the value of all possible outcomes of an experiment.

© Bushra Alhijawi

Probability Theory – Review

- For two events A and B,
  - Joint probability P(A, B) is the probability of both events occurring.
    - If events are independent, P(A, B) = P(A) * P(B).
  - Conditional probability P(A | B) is the probability of event A occurring given the previous occurrence of event B.
  - Chain rule is the fundamental relationship between joint and conditional probabilities.
    - $P(A, B) = P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$.
  - Partition rule → if an event B can be divided into a set of disjoint subcases, then the probability of B is the sum of the probabilities of the subcases.
    - $P(B) = P(A, B) + P(\bar{A}, B)$
    - $P(B) = P(A = a_1, B) + P(A = a_2, B) + P(A = a_3, B) + \ \ldots + P(A = a_N, B)$

Probability Theory – Review

- Bayes’ Rule → inverts the conditional probabilities.
  - $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
    - Prior probability P(A) → initial estimate of how likely event A is in the absence of any other information.

- Odds of an event occurring is a ratio of the probability of event occurring and it not occurring (multiplier for how probability changes).
  - $O(A) = \frac{P(A)}{P(\bar{A})} = \frac{P(A)}{1-P(A)}$

Probability Ranking Principle

- Given,

A collection of documents + User issues a query → A list of documents needs to be returned

- $R_{d,q}$ is a random variable that says whether ***d*** is relevant with respect to a given query ***q***.
  - $R_{d,q} = 1$ → the document is relevant, and $R_{d,q} = 0$ → otherwise.
- Ranking method is the core of modern IR systems → Rank by probability of relevance of the document with respect to information need.
  - $P(R = 1|d, q)$

Probability Ranking Principle (PRP)

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.”

[1960s/1970s] S. Robertson, W.S. Cooper, M.E. Maron; van Rijsbergen (1979:113); Manning & Schütze (1999:538)

Probability Ranking Principle

- In the simplest case of the PRP, there are no retrieval costs or other utility concerns that would differentially weight actions or errors.
  - You lose a point for either:
    - Returning a nonrelevant document.
    - Failing to return a relevant document.

1/0 loss

- The goal is to return the best possible results as the top-k documents for any value of k the user chooses to examine.
- The PRP then says to simply rank all documents in decreasing order of $P(R = 1|d, q)$.

Probability Ranking Principle

- If a set of retrieval results is to be returned, rather than an ordering, the Bayes Optimal Decision Rule, the decision which minimizes the risk of loss, is to simply return documents that are more likely relevant than nonrelevant:

$$d \text{ is relevant iff } P(R=1|d,q) > P(R=0|d,q)$$

- **Theorem**: Using the PRP is optimal in that it minimizes the loss (Bayes risk) under 1/0 loss
    - Provable if all probabilities are correct, etc. [e.g., Ripley 1996]

Probability Ranking Principle

- Let x represent a document in the collection.
- Let R represent the relevance of a document w.r.t. given (fixed) query.
- Let R=1 represent relevant and R=0 not relevant.

Need to find p(R=1|x) → probability that a document x is relevant.

$$P(R=1|x)=\frac{P(x|R=1)P(R=1)}{P(x)}$$

$$P(R=0|x)=\frac{P(x|R=0)P(R=0)}{P(x)}$$

$$P(R=1|x)+P(R=0|x)=1$$

- $P(R=1), P(R=0)$ - prior probability of retrieving a relevant or non-relevant document at random.
- $P(x|R=1), P(x|R=0)$- probability that if a relevant (not relevant) document is retrieved, it is x.

Probability Ranking Principle

- First, estimate how each term contributes to relevance.
  - How do other things like term frequency and document length influence your judgments about document relevance?
- Combine to find document relevance probability.
- Order documents by decreasing probability.

How do we compute all those probabilities?

Binary Independence Model

- Binary independence model (BIM) introduces two major assumptions that further simplify the computation of $P(d|q, R)$.

Binary independence model

- “Binary” = Boolean: documents are represented as binary incidence vectors of terms.
  - $\vec{d}\ (d_1, d_2, \dots, d_n)$
  - $d_i = 1$ if term *i* is present in document *d*.

- “Independence”: terms occur in documents independently.
- Different documents can be modeled as the same vector.

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.
- Using Odds and Bayes' Rule:

$$O(R|q,\vec{x}) = \frac{O(R=1|q,\vec{x})}{O(R=0|q,\vec{x})} = \frac{\frac{P(R=1|q)P(\vec{x}|R=1,q)}{P(\vec{x}|q)}}{\frac{P(R=0|q)P(\vec{x}|R=0,q)}{P(\vec{x}|q)}}$$

$$= \frac{P(R=1|q)}{P(R=0\ |q)} \cdot \frac{P(\vec{x}|R=1,q)}{P(\vec{x}|R=0,q)}$$

Binary Independence Model

$$O(R|q,\vec{x}) = \boxed{\frac{P(R=1|q)}{P(R=0\ |q)}} \cdot \boxed{\frac{P(\vec{x}|R=1,q)}{P(\vec{x}|R=0,q)}}$$

Constant for a given query Needs estimation

- Using Independence assumption:

$$\frac{P(\vec{x}|R=1,q)}{P(\vec{x}|R=0,q)} = \prod_{i=1}^{n} \frac{P(x_i|R=1,q)}{P(x_i|R=0,q)}$$

$$O(R|q,\vec{x}) = O(R|q)\,.\prod_{i=1}^{n} \frac{P(x_i|R=1,q)}{P(x_i|R=0,q)}$$

Binary Independence Model

- Since x_i is either 0 or 1:

$$O(R|q,\vec{x}) = O(R|q) \,.\prod_{x_i=1}\frac{P(x_i=1|R=1,q)}{P(x_i=1|R=0,q)}\cdot\prod_{x_i=0}\frac{P(x_i=0|R=1,q)}{P(x_i=0|R=0,q)}$$

- Let $p_i = P(x_i = 1|R = 1, q)$; $r_i = P(x_i = 1|R = 0, q)$
- Assume, for all terms not occurring in the query $(q_i = 0)$ $p_i = r_i$.

$$O(R|q,\vec{x}) = O(R|q) \,.\prod_{x_i=q_i=1}\frac{p_i}{r_i}\cdot\prod_{\substack{x_i=0\\q_i=1}}\frac{1-p_i}{1-r_i}$$

Binary Independence Model

$$O(R|q,\vec{x}) = O(R|q) \cdot \prod_{x_i=q_i=1} \frac{p_i}{r_i} \cdot \prod_{\substack{x_i=0\\ q_i=1}} \frac{1-p_i}{1-r_i}$$

All matching terms

Non-matching query terms

	Document	Relevant (R=1)	Non-relevant (R=0)
Term present	$x_i = 1$	p_i	r_i
Term absent	$x_i = 0$	$1 - p_i$	$1 - r_i$

Binary Independence Model

$$O(R|q,\vec{x}) = O(R|q) \, . \prod_{x_i=q_i=1} \frac{p_i}{r_i} \, . \prod_{\substack{x_i=0 \\ q_i=1}} \frac{1-p_i}{1-r_i}$$

$$O(R|q,\vec{x}) = O(R|q) \, . \prod_{x_i=q_i=1} \frac{p_i}{r_i} \, . \prod_{x_i=q_i=1} \left(\frac{1-r_i}{1-p_i} \, . \frac{1-p_i}{1-r_i}\right) \prod_{\substack{x_i=0 \\ q_i=1}} \frac{1-p_i}{1-r_i}$$

$$O(R|q,\vec{x}) = O(R|q) \, . \prod_{\underset{\text{All matching terms}}{x_i=q_i=1}} \frac{p_i\,(1-r_i)}{r_i(1-p_i)} \prod_{\underset{\text{All query terms}}{q_i=1}} \frac{1-p_i}{1-r_i}$$

Binary Independence Model

Constant for a given query

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

Need to be estimated for ranking

- Rank documents equally by the logarithm of this term.
  - log is a monotonic function.
- Retrieval Status Value →

$$RSV = \log \prod_{x_i=q_i=1} \frac{p_i\,(1-r_i)}{r_i(1-p_i)} = \sum_{x_i=q_i=1} \log \frac{p_i\,(1-r_i)}{r_i(1-p_i)}$$

Binary Independence Model

$$RSV = \log \prod_{x_i=q_i=1} \frac{p_i\,(1-r_i)}{r_i(1-p_i)} = \sum_{x_i=q_i=1} \log \frac{p_i\,(1-r_i)}{r_i(1-p_i)}$$

$$RSV = \sum_{x_i=q_i=1} c_i;\ c_i = \log \frac{p_i\,(1-r_i)}{r_i(1-p_i)}$$

- The c_i are log odds ratios (of contingency table a few slides back).
  - They function as the term weights in this model

How do we compute c_i ' s from our data ?

Binary Independence Model

- For each term *i*, what would these c_i numbers look like for the whole collection?

😀 Estimating RSV coefficients in theory!

- For each term *i* look at this table of document counts:

	Documents	Relevant	Non-relevant	Total
Term present	$x_i = 1$	s	$n - s$	n
Term absent	$x_i = 0$	$S - s$	$(N - n) - (S - s)$	$N - n$
Total		S	$N - S$	N

- Estimate:

$$p_i \approx \frac{s}{S}\,;\;\; r_i \approx \frac{n-s}{N-S} \quad \Rightarrow \quad c_i \approx K(N, n, S, s)\ \log \frac{s/(S-s)}{(n-s)/(N-n-S+s)}$$

For now, assume no zero terms. Remember smoothing.

Binary Independence Model

- To avoid the possibility of zeroes (such as if every or no relevant document has a particular term) it is fairly standard to add 0.5 to each of the quantities in the center 4 terms.

$$c_i \approx K(N, n, S, s)\ \log \frac{(s+0.5)/(S-s+0.5)}{(n-s+0.5)/(N-n-S+s+0.5)}$$

- IMPOTRANT → c_i is w_i and is computed only for terms that actually appear in the document *d*.

Binary Independence Model

- Scenario #1: Estimating $P(x_i|R = 1, q)$ and $P(x_i|R = 0, q)$ without relevance Judgements (ad-hoc retrieval).
  - No user-supplied relevance judgments available.
  - Query terms equally likely to appear and not to appear in relevant documents → $P(x_i|R = 1, q) = 0.5$.
  - Probability of the term *i* appearing in irrelevant documents is proportional to the number N_i of documents in the entire document collection →$P(x_i|R = 0, q) = N_i/N$.
  - Compute the relevance score for a document for the BIM without any relevance judgements:

$$rel(D, Q) = \sum_{t \in Q} \left(\frac{P(D_t|Q, R = 1)}{P(D_t|Q, R = 0)} \right) = \sum_{t \in Q} \left(\frac{0.5}{\frac{N_t}{N}} \right) = \sum_{t \in Q} \left(0.5 \frac{N}{N_t} \right)$$

Binary Independence Model – Example 1

- Given a document collection consists of the following documents:
  - d1 : “Frodo and Sam work in google”
  - d2 : “Sam left the google last week”
  - d3 : “Sam took the gift”
- The query is: “Sam work google”.
- Rank the documents based on BIM.

Binary Independence Model – Example 1

- d1 : “Frodo and Sam work in google”
- d2 : “Sam left the google last week”
- d3 : “Sam took the gift”
- The query is: “Sam work google”.

	Document 1			Document 2		Document 3
Term	sam	work	google	sam	google	sam
$P(D_t \mid Q, R = 1)$	0.5	0.5	0.5	0.5	0.5	0.5
$P(D_t \mid Q, R = 0)$	1	1/3 = 0.34	0.67	1	0.67	1
w_t	0.5	1.5	0.75	0.5	0.75	0.5
$\sum w_t$	2.75			1.25		0.5

Note: computations in this example are done without taking the logarithm

Binary Independence Model

- Scenario #2: Estimating $P(x_i|R = 1, q)$ and $P(x_i|R = 0, q)$ with relevance Judgements.
  - Let r_t be the number of documents judged as relevant that contain term t.
  - Let R be the overall number of documents judged as relevant.
- In this setting, estimate the term-relevance probabilities as follows:
  - $P(x_i|R = 1, q) = (r_t+0.5)/(R + 1)$
  - $P(x_i|R = 0, q) = (N_t - r_t + 0.5)/(N - R + 1)$
- Compute the relevance score for a document for the BIM with any relevance judgements:

$$rel(D, Q) = \sum_{t \in Q} \left(\frac{P(D_t|Q, R = 1)}{P(D_t|Q, R = 0)} \right) = \sum_{t \in Q} \left(\frac{(r_t + 0.5) \, . \, (N - R + 1)}{(R + 1) \, . \, (N_t - r_t + 0.5)} \right)$$

Binary Independence Model – Example 1

- Given a collection contains $\mathrm{N} = 30$ documents, including:
  - d1 : “Frodo and Sam work in google”
  - d2 : “Sam left the google last week”
  - d3 : “Sam took the gift”
- The query is: “Sam work google”.
- User has indicated $\mathrm{R} = 6$ relevant documents for this query.
- Query terms: t1 = “Sam”, t2 = “work”, t3 = “google”
- Document frequencies of query terms in relevant documents and overall collection are given as follows:
  - $r_{t1} = 3;\ \ N_{t1} = 15$
  - $r_{t2} = 4;\ \ N_{t2} = 16$
  - $r_{t3} = 2;\ \ N_{t3} = 14$

Binary Independence Model – Example 1

- d1 : “Frodo and Sam work in google”
- d2 : “Sam left the google last week”
- d3 : “Sam took the gift”
- The query is: “Sam work google”.

	Document 1			Document 2		Document 3
Term	sam	work	google	sam	google	sam
$P(D_t \mid Q, R = 1)$	0.5	0.64	0.36	0.5	0.36	0.5
$P(D_t \mid Q, R = 0)$	0.5	0.5	0.5	0.5	0.5	0.5
w_t	1	1.28	0.72	1	0.72	0.5
$\sum w_t$	3			1.72		0.5

Note: computations in this example are done without taking the logarithm

Any Question

Princess Sumaya University for Technology
جامعـــة الأميـرة سميّـة للتكنولوجيا

Excellence in Education, since 1991 التميّـز في التعليـم منـذ عام 1991

www.psut.edu.jo

Call: (+962) 6-5359 949
Fax: (+962) 6-5347 295
Email: info@psut.edu.jo

Princess Sumaya University for Technology
Amman 11941 Jordan
P.o.Box 1438 Al-Jubaiha