Information Retrieval

Dhruv Gupta

dhruv.gupta@ntnu.no

13-September-2022

NTNU | Norwegian University of Science and Technology

Announcements

- Assignment 1: available and due on 22.September.2022.
- Reference Group: volunteers needed for feedback regarding course.
    - Interested? Please contact me by email!

- Text and diagrams of some slides are based on the material from the book: Baeza-Yates and Ribeiro-Neto, “Modern Information Retrieval", Second Edition. Pearson Education Limited, 2011.

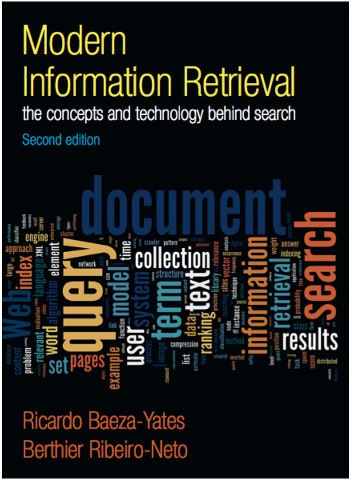
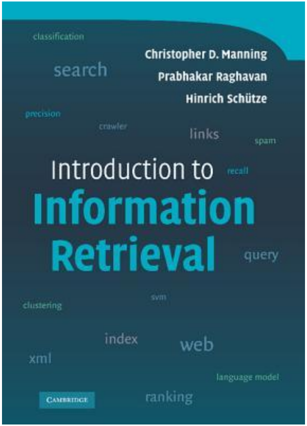


Image Credit: `http://grupoweb.upf.es/mir2ed/`

References

- Text and diagrams of some slides are based on the material from the book: Manning et al., “Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.
- Some slides for the Evaluation topic are adapted from Hinrich Schütze’s lectures at LMU.[1]



[1] https://www.cis.lmu.de/~hs/teach/14s/ir/

Image Credit: https://www.goodreads.com/book/show/3278309-introduction-to-information-retrieval

Recap — 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."

— van Rijsbergen, 1979.

All from: Manning et al., "Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Recap — The Binary Independence Model

- Binary Independence Model Assumption 2: $\forall w_j \notin q, p_{iR} = q_{iR}$.
- Converting log products into sums of logs, we have

$$\mathrm{sim}(d_i, q) \sim \sum_{w_j \in q \wedge w_j \in d_i} \log\left[\frac{p_{iR}}{1 - p_{iR}}\right] + \log\left[\frac{1 - q_{iR}}{p_{iR}}\right].$$

- The above formula is a key expression for ranking computation in the probabilistic model.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Recap — Ranking Formula

- In the previous formula, we are still dependent on an estimation of the relevant docs for the query.
- For handling small values of r_j, we add 0.5 to each of the terms in the formula above, which changes $\mathrm{sim}(d_i, q)$ into

$$\mathrm{sim}(d_i, q) \sim \sum_{w_j \in q \wedge w_j \in d_i} \log\left(\frac{r_j + 0.5}{R - r_j + 0.5} \cdot \frac{N - n_j - R + r_j + 0.5}{n_j - r_j + 0.5}\right).$$

- This formula is considered as the classic ranking equation for the probabilistic model and is known as the Robertson-Sparck Jones Equation.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Recap — The Probabilistic Model

- The probabilistic ranking formula:

$$\text{sim}(d_i, q) \sim \sum_{w_j \in q \wedge w_j \in d_i} \log\left[\frac{N - n_j}{n_j}\right].$$

- To avoid problems with $D = 1$ and $D_j = 0$:

$$p_{jR} = \frac{D_j + 0.5}{D + 1} \text{ and } q_{jR} = \frac{n_j - D_j + 0.5}{N - D + 1}.$$

- Also,

$$p_{jR} = \frac{D_j + \frac{n_j}{N}}{D + 1} \text{ and } q_{jR} = \frac{n_j - D_j + \frac{n_j}{N}}{N - D + 1}.$$

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Recap — BM25 Ranking Formula

- BM25: combination of the BM11 and BM15.
- The motivation was to combine the BM11 and BM15 term frequency factors as follows.

$$\mathcal{B}_{i,j} = \frac{(K_1 + 1) \cdot \mathrm{tf}_{i,j}}{K_1 \cdot \left[(1 - b) + b \cdot \frac{\mathrm{len}(d_i)}{\mathrm{avg\_doclen}}\right] + \mathrm{tf}_{i,j}}.$$

- where, b is a constant with values in the interval $[0, 1]$.
  - If $b = 0$, it reduces to the BM15 term frequency factor.
  - If $b = 1$, it reduces to the BM11 term frequency factor.
  - For values of $b \in (0, 1)$, the equation provides a combination of BM11 and BM15.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Recap — BM25 Ranking Formula

- The ranking equation for the BM25 model can then be written as:

$$\text{sim}_{\text{BM25}}(d_i, q) \sim \sum_{w_j \in q \wedge w_j \in d_i} \mathcal{B}_{i,j} \cdot \log\left[\frac{N - n_j + 0.5}{n_j + 0.5}\right]$$

- where, K_1 and b are empirical constants.
  - $K_1 = 1$ works well with real collections.
  - b should be kept closer to 1 to emphasize the document length normalization effect present in the BM11 formula.
  - For instance, $b = 0.75$ is a reasonable assumption.
  - Constants values can be fine tuned for particular collections through proper experimentation.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Recap — Statistical Language Models in IR

- Each document is treated as (the basis for) a language model.
- Given a query q.
- Rank documents based on $P(d|q)$.

$$P(d|q) = \frac{P(q|d) \cdot P(d)}{P(q)} \propto P(q|d) \cdot P(d)$$

- $P(q)$ is the same for all documents, so ignore.
- $P(d)$ is the prior – often treated as the same for all d.
  - But we can give a higher prior to “high-quality” documents, e.g., those with high PageRank.
- $P(q|d)$ is the probability of q given d.
- Under the assumptions we made, ranking documents according to $P(q|d) \cdot P(d)$ or $P(d|q)$ is considered equivalent.

All from: Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

Recap — Jelinek-Mercer Smoothing

- Jelinek-Mercer Smoothing:

$$P(q|d) \propto \prod_{1 \leq k \leq |q|} \lambda \cdot P(t_k|M_d) + (1 - \lambda) \cdot P(t_k|M_c)$$

- What we model: the user has a document in mind and generates the query from this document.
- $P(q|d)$ is the probability that the document that the user had in mind was in fact this one.

All from: Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

Recap — Dirichlet Smoothing

- Dirichlet Smoothing:

$$P(t|d) = \frac{\mathrm{tf}_{d,t} + \mu \cdot P(t|M_c)}{|d| + \mu}$$

- The background distribution $P(t|M_c)$ is the prior for $P(t|d)$.
- Intuition: before having seen any part of the document we start with the background distribution as our estimate.
- As we read the document and count terms we update the background distribution.
- The weighting factor μ determines how strong an effect the prior has.

All from: Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

Exercise 1 — Probabilistic Ranking Functions

- We considered three key qualities that a ranking function for an IR system should contain.
- What are the qualities present for the following ranking formula:

$$\mathrm{sim}(d, q) \sim \sum_{t \in q} \frac{tf_{d,t} \cdot (k_1 + 1)}{k_1 + tf_{d,t}} \cdot \log\left[\frac{N}{n_t}\right]$$

- where,
  - tf indicates term frequency of term t in document d.
  - k_1 is a constant > 0.
  - N is the number of documents in the collection ($= |\mathcal{D}|$).
  - n_t is the number of documents containing term t.

All from: Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

Exercise 2 — Compute Ranking using Language Models

- Jelinek-Mercer Smoothing:

$$P(q|d) \propto \prod_{1 \le k \le |q|} \lambda \cdot P(t_k|M_d) + (1-\lambda) \cdot P(t_k|M_c)$$

- Dirichlet Smoothing:

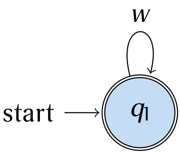
$$P(q|d) \propto \prod_{1 \le k \le |q|} \frac{\mathrm{tf}_{d,t_k} + \mu \cdot P(t_k|M_c)}{|d| + \mu}$$

- Collection: d_1 and d_2.
- d_1: *Xerox reports a profit but revenue is down*.
- d_2: *Lucene narrows quarter loss but revenue decreases further*.
- q: *revenue down*.
- Compute ranking using Jelinek-Mercer smoothing with $\lambda = 1/2$.
- Compute ranking using Dirichlet smoothing with $\mu = 8$.

All from: Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

Statistical Language Models — Implementation Issue

- This is a one-state probabilistic finite-state automaton – a unigram language model – and the state emission distribution for its one state q_1.
- STOP is not a word, but a special symbol indicating that the automaton stops.



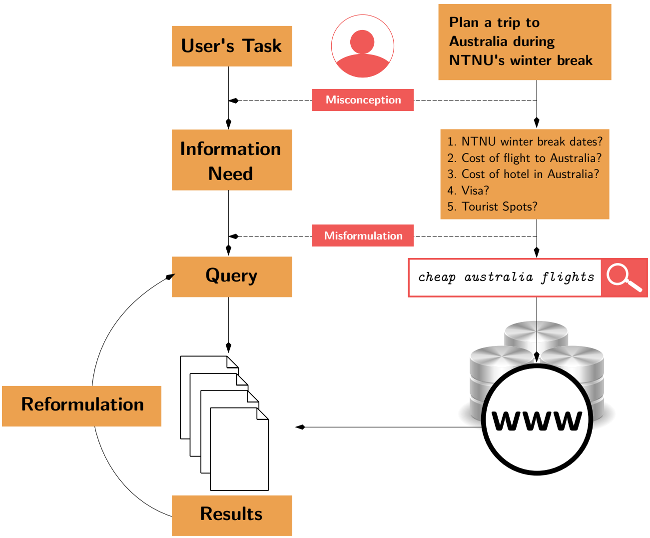
w	$P(w\|q_1)$	w	$P(w\|q_1)$
STOP	0.2	toad	0.01
the	0.2	said	0.03
a	0.1	likes	0.02
frog	0.01	that	0.04
		...	...

frog said that toad likes frog STOP

$$P(\text{string}) = 0.01 \cdot 0.03 \cdot 0.04 \cdot 0.01 \cdot 0.02 \cdot 0.01 \cdot 0.2 = 0.000000000048$$

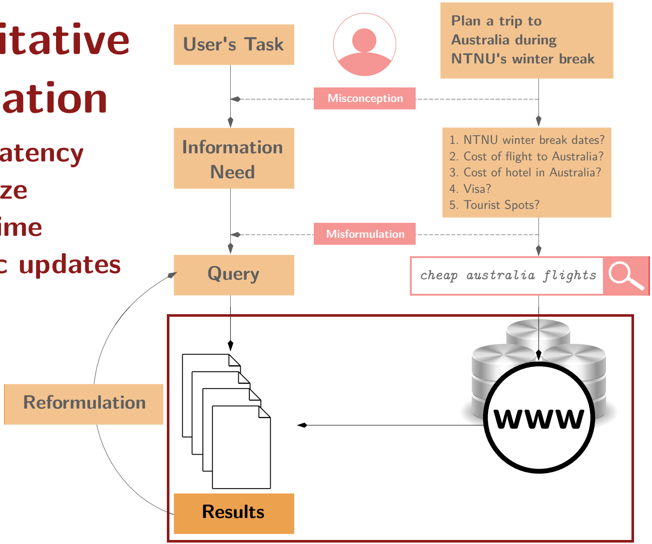
All from: Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

Information Retrieval — Evaluation



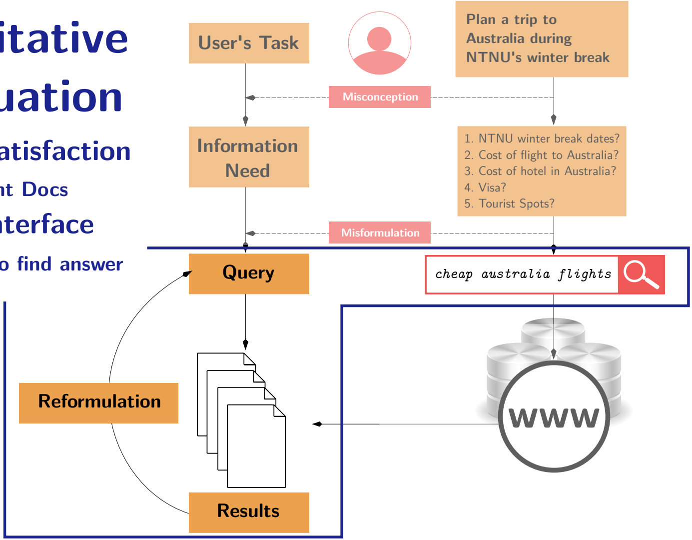
Quantitative Evaluation

- Search latency
- Index Size
- Index Time
- Dynamic updates



Qualitative Evaluation

- User Satisfaction
  - Relevant Docs
- User Interface
  - Time to find answer



Relevance — Query vs. Information Need

- User satisfaction is equated with the relevance of search results to the query.
- Information Need i: *I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.*
- This is an information need, not a query.
- Query q: *red wine white wine heart attack*.
- Consider document d': *At the heart of his speech was an attack on the wine industry lobby for downplaying the role of red and white wine in drunk driving.*
- d' is an good match for query q.
- d' is not relevant to the information need i.

Manning et al., "Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Schütze et al.: `https://www.cis.lmu.de/~hs/teach/14s/ir/`.

Relevance — Query vs. Information Need

- User satisfaction can only be measured by relevance to an information need, not by relevance to queries.
- Note on terminology: query-document relevance judgments are sometimes equated to information-need-document relevance judgments.

Manning et al., "Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Schütze et al.: `https://www.cis.lmu.de/~hs/teach/14s/ir/`.

The Cranfield Paradigm

- Evaluation of IR systems is the result of early experimentation initiated in the 50's by Cyril Cleverdon.
- The insights derived from these experiments provide a foundation for the evaluation of IR systems.
- Back in 1952, Cleverdon took notice of a new indexing system called Uniterm, proposed by Mortimer Taube.
- Cleverdon thought it appealing and with Bob Thorne, a colleague, did a small test:
  - He manually indexed 200 documents using Uniterm and asked Thorne to run some queries.
  - This experiment put Cleverdon on a life trajectory of reliance on experimentation for evaluating indexing systems.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Cranfield Paradigm

- Cleverdon obtained a grant from the National Science Foundation to compare distinct indexing systems.
- These experiments provided interesting insights, that culminated in the modern metrics of precision and recall.
- Recall ratio: the fraction of relevant documents retrieved.
- Precision ratio: the fraction of documents retrieved that are relevant.
- For instance, it became clear that, in practical situations, the majority of searches does not require high recall.
- Instead, the vast majority of the users require just a few relevant answers.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Cranfield Paradigm

- The next step was to devise a set of experiments that would allow evaluating each indexing system in isolation more thoroughly.
- The result was a test reference collection composed of documents, queries, and relevance judgements.
- It became known as the Cranfield-2 collection.
- The reference collection allows using the same set of documents and queries to evaluate different ranking systems.
- The uniformity of this setup allows quick evaluation of new ranking functions.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Cranfield Paradigm

- The next step was to devise a set of experiments that would allow evaluating each indexing system in isolation more thoroughly.
- The result was a test reference collection (Cranfield-2 collection) composed of:
  1. Documents,
  2. Queries, and
  3. Relevance Judgements.
- The reference collection allows using the same set of documents and queries to evaluate different ranking systems.
- The uniformity of this setup allows quick evaluation of new ranking functions.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Reference Collections

- Reference collections, which are based on the foundations established by the Cranfield experiments, constitute the most used evaluation method in IR.
- A reference collection is composed of:
  1. A set $\mathcal{D}$ of pre-selected documents.
  2. A set $\mathcal{I}$ of information need descriptions used for testing.
  3. A set of relevance judgements associated with each pair $[i, d]$, where, $i \in \mathcal{I}$ and $d \in \mathcal{D}$.
- The relevance judgement has a value of:
  - 0 if document d is non-relevant to i.
  - 1 if document d is relevant to i.
- These judgements are produced by human specialists.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Precision and Recall

- Precision (P) is the fraction of retrieved documents that are relevant:

$$Precision = \frac{\#\text{relevant documents retrieved}}{\#\text{retrieved documents}} = P(\text{relevant}|\text{retrieved}).$$

- Recall (R) is the fraction of relevant documents that are retrieved:

$$Recall = \frac{\#\text{relevant documents retrieved}}{\#\text{relevant documents}} = P(\text{retrieved}|\text{relevant}).$$

Manning et al., "Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Schütze et al.: `https://www.cis.lmu.de/~hs/teach/14s/ir/`.

Precision and Recall

	Relevant	**Non-Relevant**
Retrieved	TP (True Positive)	FP (False Positive)
Not Retrieved	FN (False Negative)	TN (True Negative)

$$\text{Precision} = P = \frac{TP}{TP + FP}$$

$$\text{Recall} = R = \frac{TP}{TP + FN}$$

Manning et al., "Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Schütze et al.: `https://www.cis.lmu.de/~hs/teach/14s/ir/`.

Precision-Recall Trade-Off

- You can increase recall by returning more documents.
- Recall is a non-decreasing function of the number of documents retrieved.
- A system that returns all docs has 100% recall!
- The converse is also true (usually): It's easy to get high precision for very low recall.

Manning et al., "Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Schütze et al.: `https://www.cis.lmu.de/~hs/teach/14s/ir/`.

Combining Precision and Recall: F-measure

- F-measure allows us to trade off precision against recall.

$$F = \frac{1}{\alpha \cdot \frac{1}{P} + (1-\alpha) \cdot \frac{1}{R}} = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R}, \qquad \text{where } \beta^2 = \frac{1-\alpha}{\alpha}.$$

- $\alpha \in [0, 1]$ and therefore $\beta^2 \in [0, \infty]$.
- Usually used: balanced F-measure with $\alpha = 0.5$ or $\beta = 1$. This correspond to the harmonic mean of precision and recall.

$$\frac{1}{F} = \frac{1}{2} \cdot \left(\frac{1}{P} + \frac{1}{R}\right) = \frac{P + R}{2 \cdot P \cdot R}.$$

- The F-measure is derived from the E-measure with the following equivalence:

$$E = 1 - F = 1 - \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R}.$$

Manning et al., “Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Schütze et al.: https://www.cis.lmu.de/~hs/teach/14s/ir/.

F-measure

- The parameter β is specified by the user and reflects the relative importance of recall and precision.
- If $\beta = 0$:
  - $F = P$.
  - Low values of β make F-measure a function of precision.
- If $\beta \rightarrow \infty$:
  - $\lim_{\beta \rightarrow \infty} F = R$.
  - High values of β make F-measure a function of recall.
- For $\beta = 1$ we get a harmonic mean of precision and recall.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

F-measure

- Why harmonic mean for combining precision and recall? Why not arithmetic mean?
- Recall: if a system returns the entire document collection for each query then recall is 100%.
- $P = 0, R = 1 \rightarrow$ Arithmetic Mean $= 0.5$.
- Harmonic Mean (HM) $\leq$ Geometric Mean (GM) $\leq$ Arithmetic Mean (AM).
- When values of two numbers differ greatly, the HM is closer to the minimum than their AM.

Manning et al., "Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Schütze et al.: https://www.cis.lmu.de/~hs/teach/14s/ir/.

Precision and Recall

	Relevant	**Non-Relevant**	
Retrieved	20 (TP)	40 (FP)	60
Not Retrieved	60 (FN)	1,000,000 (TN)	1,000,060
	80	1,000,040	1,000,120

$$\text{Precision} = P = \frac{TP}{TP + FP} = \frac{20}{20 + 40} = \frac{1}{3}$$
$$\text{Recall} = R = \frac{TP}{TP + FN} = \frac{20}{20 + 60} = \frac{1}{4}$$
$$\text{F-Measure} = \mathrm{F}_1 = \frac{2 \cdot P \cdot R}{P + R} = \frac{2 \cdot \frac{1}{3} \cdot \frac{1}{4}}{\frac{1}{3} + \frac{1}{4}} = \frac{2}{7}$$

Manning et al., "Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Schütze et al.: `https://www.cis.lmu.de/~hs/teach/14s/ir/`.

Accuracy

	Relevant	**Non-Relevant**
Retrieved	TP (True Positive)	FP (False Positive)
Not Retrieved	FN (False Negative)	TN (True Negative)

- Accuracy: fraction of decisions (relevant/non-relevant) that are correct.
- In terms of the contingency table above:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}.$$

- But there is a problem!

Manning et al., "Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Schütze et al.: `https://www.cis.lmu.de/~hs/teach/14s/ir/`.

Exercise

	Relevant	**Non-Relevant**
Retrieved	18 (TP)	2 (FP)
Not Retrieved	82 (FN)	1,000,000,000 (TN)

- Compute precision, recall, F_1, and Accuracy.

$$\text{Precision} = P = \frac{TP}{TP + FP} = ?$$

$$\text{Recall} = R = \frac{TP}{TP + FN} = ?$$

$$\text{F-Measure} = F_1 = \frac{2 \cdot P \cdot R}{P + R} = ?$$

$$\text{Accuracy} = A = \frac{TP + TN}{TP + FP + FN + TN} = ?$$

Manning et al., "Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Schütze et al.: `https://www.cis.lmu.de/~hs/teach/14s/ir/`.

Accuracy

- Accuracy: is a bad for measuring performance as the distribution of relevant and non-relevant documents per query is very skewed.
- A ranking function will get high accuracy even if it labels zero relevant documents!
- Information Retrieval algorithms can tolerate a degree of non-relevant results in the answer set.
- Better to optimize for Precision, Recall, and F_1 measure were number of relevant results in the answer set can be assessed more reliably.

Manning et al., "Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Schütze et al.: `https://www.cis.lmu.de/~hs/teach/14s/ir/`.

Precision and Recall

- The definition of precision and recall assumes that all docs in the answer set have been examined.
- However, the user is not usually presented with all docs in the answer set at once.
- Instead, the user sees a ranked set of documents and examines them starting from the top.
- Thus, precision and recall vary as the user examines the document list.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Precision and Recall

- Consider a reference collection and a set of test queries.
- Let R_{q_1} be the set of relevant docs for a query q_1:

$$R_{q_1} = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}.$$

- Consider a new IR algorithm that yields the following answer to q_1 (relevant docs are marked with a bullet):

01. d_{123} •	06. d_9 •	11. d_{38}
02. d_{84}	07. d_{511}	12. d_{48}
03. d_{56} •	08. d_{129}	13. d_{250}
04. d_6	09. d_{187}	14. d_{113}
05. d_8	10. d_{25} •	15. d_3 •

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Precision and Recall

- If we examine this ranking, we observe that:
- The document d_{123}, ranked as number 1, is relevant.
  - This document corresponds to 10% of all relevant documents.
  - Thus, we say that we have a precision of 100% at 10% recall.

01. d_{123} •	06. d_9 •	11. d_{38}
02. d_{84}	07. d_{511}	12. d_{48}
03. d_{56} •	08. d_{129}	13. d_{250}
04. d_6	09. d_{187}	14. d_{113}
05. d_8	10. d_{25} •	15. d_3 •

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Precision and Recall

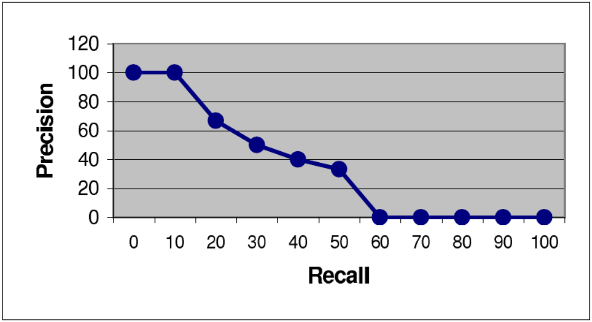
- If we examine this ranking, we observe that:
- The document d_{56}, ranked as number 3, is relevant.
  - At this point, two documents out of three are relevant, and two of the ten relevant documents have been seen.
  - Thus, we say that we have a precision of $66.\bar{6}\%$ at 20% recall.

01. d_{123} •	06. d_9 •	11. d_{38}
02. d_{84}	07. d_{511}	12. d_{48}
03. d_{56} •	08. d_{129}	13. d_{250}
04. d_6	09. d_{187}	14. d_{113}
05. d_8	10. d_{25} •	15. d_3 •

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Precision and Recall

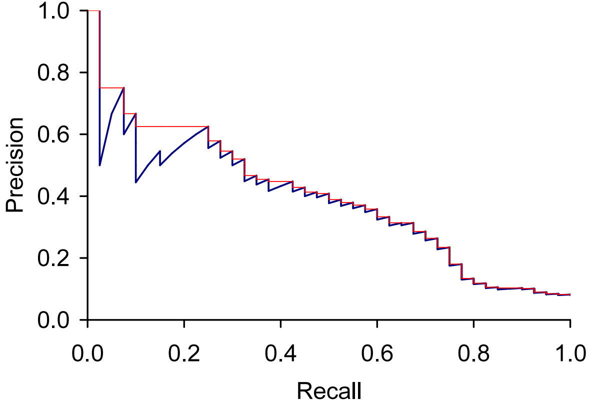
- If we proceed with our examination of the ranking generated, we can plot a curve of precision versus recall as follows:



Recall	Precision
0	100
10	100
20	66.6
30	50
40	40
50	33.3
60	0
70	0
80	0
90	0
100	0

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Precision and Recall Curve



All from: Manning et al., “Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Precision and Recall

- Consider now a second query q_2
  whose set of relevant answers is given by:

$$R_{q_2} = \{d_3, d_{56}, d_{129}\}.$$

- The previous IR algorithm processes the query q_2 and returns a ranking, as follows:

01. d_{425}	06. d_{615}	11. d_{193}
02. d_{87}	07. d_{512}	12. d_{715}
03. d_{56} •	08. d_{129} •	13. d_{810}
04. d_{32}	09. d_4	14. d_5
05. d_{124}	10. d_{130}	15. d_3 •

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Precision and Recall

- If we examine this ranking, we observe:
- The first relevant document is d_{56}.
  - It provides a recall and precision levels equal to $33.\overline{3}\%$.

01. d_{425}	06. d_{615}	11. d_{193}
02. d_{87}	07. d_{512}	12. d_{715}
03. d_{56} •	08. d_{129} •	13. d_{810}
04. d_{32}	09. d_{4}	14. d_{5}
05. d_{124}	10. d_{130}	15. d_{3} •

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Precision and Recall

- If we examine this ranking, we observe:
- The second relevant document is d_{129}.
  - It provides a recall level of $66.\bar{6}\%$ (with precision equal to 25%).

01. d_{425}	06. d_{615}	11. d_{193}
02. d_{87}	07. d_{512}	12. d_{715}
03. d_{56} •	08. d_{129} •	13. d_{810}
04. d_{32}	09. d_{4}	14. d_{5}
05. d_{124}	10. d_{130}	15. d_{3} •

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Precision and Recall

- If we examine this ranking, we observe:
- The third relevant document is d_3.
  - It provides a recall level of 100% (with precision equal to 20%).

01. d_{425}	06. d_{615}	11. d_{193}
02. d_{87}	07. d_{512}	12. d_{715}
03. d_{56} •	08. d_{129} •	13. d_{810}
04. d_{32}	09. d_{4}	14. d_{5}
05. d_{124}	10. d_{130}	15. d_{3} •

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

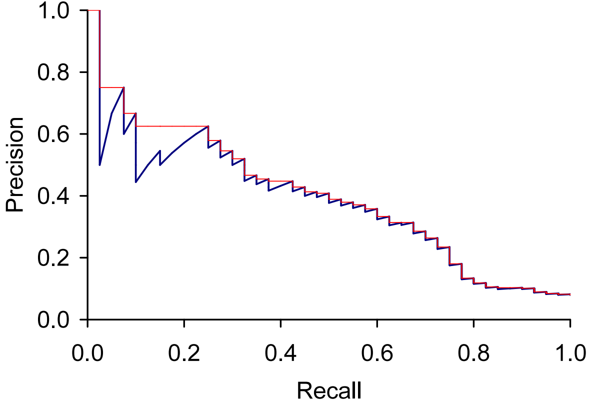
- The precision figures at the 11 standard recall levels are interpolated as follows.
- Let r_j, where $j \in \{0, 1, 2, \ldots, 10\}$, be a reference to the j–th standard recall level. Then,

$$P(r_j) = \max_{\forall r | r_j \leq r} P(r)$$

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Precision and Recall Curve — Interpolation

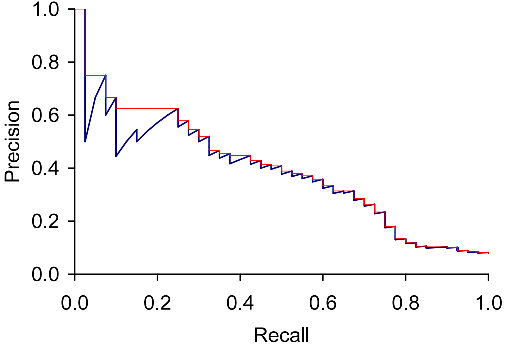
- The saw-tooth shape of the curve can be "smoothed" by using interpolation.



All from: Manning et al., "Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Precision and Recall Curve — Interpolation

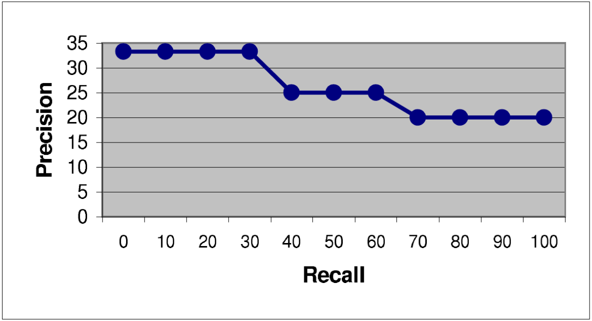
- Another justification for interpolation: users are willing to look at a few more documents if it would increase the percentage of the viewed set that were relevant (that is, if the precision of the larger set is higher).
- This way precision at recall level of 0 is defined.



All from: Manning et al., "Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Precision and Recall

- In our last example, this interpolation rule yields the precision and recall figures illustrated below:



Recall	Precision
0	33.3
10	33.3
20	33.3
30	33.3
40	25
50	25
60	25
70	20
80	20
90	20
100	20

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Precision and Recall

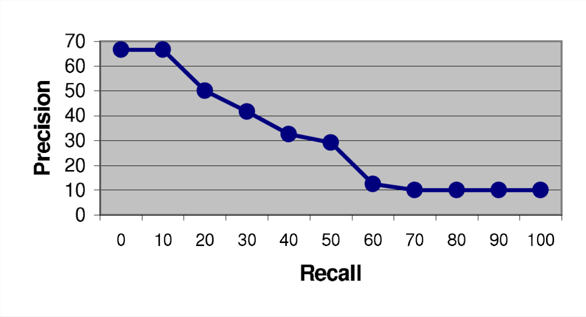
- In the examples above, the precision and recall figures have been computed for single queries.
- Usually, however, retrieval algorithms are evaluated by running them for several distinct test queries.
- To evaluate the retrieval performance for N_q queries, we average the precision at each recall level as follows:

$$\bar{P}(r_j) = \sum_{i=1}^{N_q} \frac{P_i(r_j)}{N_q}.$$

- where,
  - $\bar{P}(r_j)$ is the average precision at recall level r_j.
  - $P(r_j)$ is the precision at recall level r_j for the i–th query.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

- To illustrate, the figure below illustrates precision-recall figures averaged over queries q_1 and q_2:

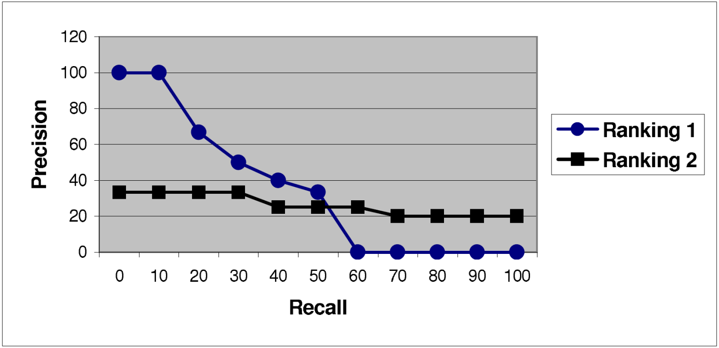


Recall	Precision
0	66.6
10	66.6
20	49.9
30	41.6
40	32.5
50	29.1
60	12.5
70	10
80	10
90	10
100	10

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Precision and Recall

- Average precision-recall curves are normally used to compare the performance of distinct IR algorithms.
- The figure below illustrates average precision-recall curves for two distinct retrieval algorithms.



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Precision and Recall Appropriateness

- Precision and recall have been extensively used to evaluate the retrieval performance of IR algorithms.
- However, a more careful reflection reveals problems with these two measures:
  - First, the proper estimation of maximum recall for a query requires detailed knowledge of all the documents in the collection.
  - Second, in many situations the use of a single measure could be more appropriate.
  - Third, recall and precision measure the effectiveness over a set of queries processed in batch mode.
  - Fourth, for systems which require a weak ordering though, recall and precision might be inadequate.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Single Value Summaries

- Average precision-recall curves constitute standard evaluation metrics for information retrieval systems.
- However, there are situations in which we would like to evaluate retrieval performance over individual queries.
- The reasons are two-fold:
  - First, averaging precision over many queries might disguise important anomalies in the retrieval algorithms under study.
  - Second, we might be interested in investigating whether a algorithm outperforms the other for each query.
- In these situations, a single precision value can be used.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Precision at k: $P@K$

- In the case of Web search engines, the majority of searches does not require high recall.
- Higher the number of relevant documents at the top of the ranking, more positive is the impression of the users.
- Precision at 5 ($P@5$) and at 10 ($P@10$) measure the precision when 5 or 10 documents have been seen.
- These metrics assess whether the users are getting relevant documents at the top of the ranking or not.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Precision at k: $P@K$

- To exemplify, consider again the ranking for the example query q_1 we have been using.
- For this query, we have $P@5 = 40\%$ and $P@10 = 40\%$.
- Further, we can compute $P@5$ and $P@10$ averaged over a sample of 100 queries, for instance.
- These metrics provide an early assessment of which algorithm might be preferable in the eyes of the users.

01. d_{123} •	06. d_9 •	11. d_{38}
02. d_{84}	07. d_{511}	12. d_{48}
03. d_{56} •	08. d_{129}	13. d_{250}
04. d_6	09. d_{187}	14. d_{113}
05. d_8	10. d_{25} •	15. d_3 •

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Mean Average Precision: MAP

- The idea here is to average the precision figures obtained after each new relevant document is observed.
- For relevant documents not retrieved, the precision is set to 0.
- MAP_i: the mean value precision for query q_i is:

$$\text{MAP}_i = \frac{1}{|R_i|} \cdot \sum_{k=1}^{|R_i|} P(R_i[k]).$$

- where, R_i is the set of relevant documents for query q_i.
- where, $P(R_i[k])$ is the precision when the $R_i[k]$ document is observed in the ranking of q_i.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Mean Average Precision: MAP

- MAP: the mean average precision over a set of queries, is defined as:

$$\text{MAP} = \frac{1}{N_q} \cdot \sum_{i=1}^{N_q} \text{MAP}_i.$$

- where, N_q is the total number of queries.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Mean Average Precision: MAP

- To illustrate, consider again the ranked list of documents returned for the example query q_1.

$$R_{q_1} = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}.$$

01. d_{123} •	06. d_9 •	11. d_{38}
02. d_{84}	07. d_{511}	12. d_{48}
03. d_{56} •	08. d_{129}	13. d_{250}
04. d_6	09. d_{187}	14. d_{113}
05. d_8	10. d_{25} •	15. d_3 •

$$\text{MAP}_1 = \frac{1 + 0.66 + 0.5 + 0.4 + 0.33 + 0 + 0 + 0 + 0 + 0}{10} = 0.28.$$

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Mean Average Precision: MAP

- To illustrate, consider again the ranked list of documents returned for the example query q_2.

$$R_{q_2} = \{d_3, d_{56}, d_{129}\}.$$

01. d_{425}	06. d_{615}	11. d_{193}
02. d_{87}	07. d_{512}	12. d_{715}
03. d_{56} •	08. d_{129} •	13. d_{810}
04. d_{32}	09. d_{4}	14. d_{5}
05. d_{124}	10. d_{130}	15. d_{3} •

$$\mathrm{MAP}_2 = \frac{0.33 + 0.25 + 0.20}{3} = 0.26,$$

$$\mathrm{MAP} = \frac{\mathrm{MAP}_1 + \mathrm{MAP}_2}{2} = \frac{0.28 + 0.26}{2} = 0.27.$$

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

R-Precision

- Let R be the total number of relevant docs for a given query.
- The idea here is to compute the precision at the R–th position in the ranking.
- Example: consider query q_1,
  - The R value is 10 and there are 4 relevant documents among the top-10 documents in the ranking.
  - Thus, the R-Precision value for q_1 is $\frac{4}{10} = 0.4$.
- Example: consider query q_2,
  - The R value is 3 and there is 1 relevant document among the top-3 documents in the ranking.
  - Thus, the R-Precision value for q_2 is $\frac{1}{3} = 0.\overline{3}$.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

R-Precision

- The R-precision measure is a useful for observing the behavior of an algorithm for individual queries.
- Additionally, one can also compute an average R-precision figure over a set of queries.
  - However, using a single number to evaluate a algorithm over several queries might be quite imprecise.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Mean Reciprocal Rank: MRR

- MRR is a good metric for those cases in which we are interested in the first correct answer such as:
    - Question-Answering (QA) systems.
    - Search engine queries that look for specific sites:
        - URL Queries.
        - Homepage queries.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Mean Reciprocal Rank: MRR

- Let,
  - $\mathcal{R}_i$: ranking relative to a query q_i.
  - $S_{\text{correct}}(\mathcal{R}_i)$: position of the first correct answer in $\mathcal{R}_i$.
  - S_h: threshold for ranking position.
- Then, the reciprocal rank $\text{RR}(\mathcal{R}_i)$ for query q_i is given by:

$$\text{RR}(\mathcal{R}_i) = \begin{cases} \frac{1}{S_{\text{correct}}(\mathcal{R}_i)}, & \text{if } S_{\text{correct}}(\mathcal{R}_i) \leq S_h \\ 0, & \text{otherwise} \end{cases}$$

- The mean reciprocal rank (MRR) for a set Q of N_q queries is given by:

$$\text{MRR}(Q) = \frac{1}{N_q} \cdot \sum_{i}^{N_q} \text{RR}(\mathcal{R}_i).$$

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Mean Reciprocal Rank: MRR

- To illustrate, consider again the ranked list of documents returned for the example query q_1.

01. d_{123} •	06. d_9 •	11. d_{38}
02. d_{84}	07. d_{511}	12. d_{48}
03. d_{56} •	08. d_{129}	13. d_{250}
04. d_6	09. d_{187}	14. d_{113}
05. d_8	10. d_{25} •	15. d_3 •

$$\mathrm{RR}_1 = \frac{1}{1} = 1.$$

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Mean Reciprocal Rank: MRR

- To illustrate, consider again the ranked list of documents returned for the example query q_2.

01. d_{425}	06. d_{615}	11. d_{193}
02. d_{87}	07. d_{512}	12. d_{715}
03. d_{56} •	08. d_{129} •	13. d_{810}
04. d_{32}	09. d_{4}	14. d_{5}
05. d_{124}	10. d_{130}	15. d_{3} •

$$\mathrm{RR}_2 = \frac{1}{3} = 0.\overline{3},$$

$$\mathrm{MRR} = \frac{\mathrm{RR}_1 + \mathrm{RR}_2}{2} = \frac{1 + \frac{1}{3}}{2} = \frac{2}{3} = 0.\overline{6}.$$

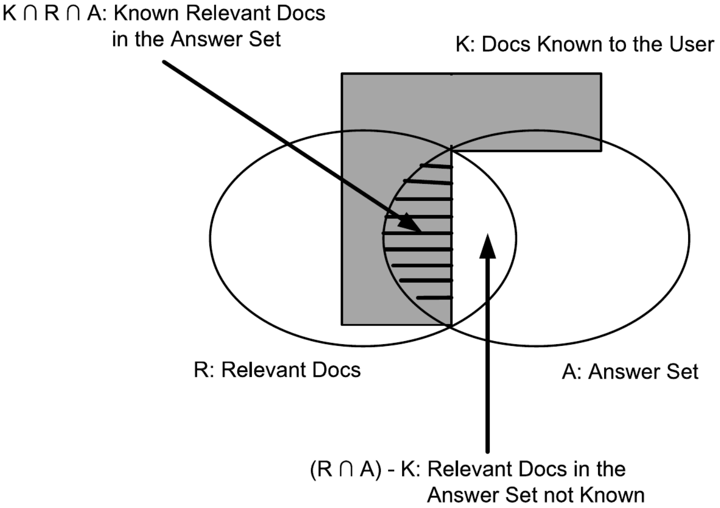
Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

User-Oriented Measures

- Recall and precision assume that the set of relevant docs for a query is independent of the users.
- However, different users might have different relevance interpretations.
- To cope with this problem, user-oriented measures have been proposed.
- As before,
  - Consider a reference collection, an information request I, and a retrieval algorithm to be evaluated.
  - with regard to I, let R be the set of relevant documents and A be the set of answers retrieved.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

User-Oriented Measures



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

User-Oriented Measures

- The coverage ratio is defined as the fraction of the documents known and relevant that are in the answer set, that is:

$$\text{coverage} = \frac{|K \cap R \cap A|}{|K \cap R|}.$$

- A high coverage indicates that the system has found most of the relevant docs the user expected to see.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

User-Oriented Measures

- The novelty ratio is defined as the fraction of the relevant documents in the answer set that are not known to the user, that is:

$$\text{novelty} = \frac{|(R \cap K) - A|}{|R \cap A|}.$$

- A high novelty indicates that the system is revealing many new relevant docs which were unknown.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

User-Oriented Measures: Additional Measures

- Relative Recall: ratio between the number of relevant docs found and the number of relevant docs the user expected to find.
- Recall Effort: ratio between the number of relevant docs the user expected to find and the number of documents examined in an attempt to find the expected relevant documents.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.