Modern Information Retrieval

Evaluation in information retrieval

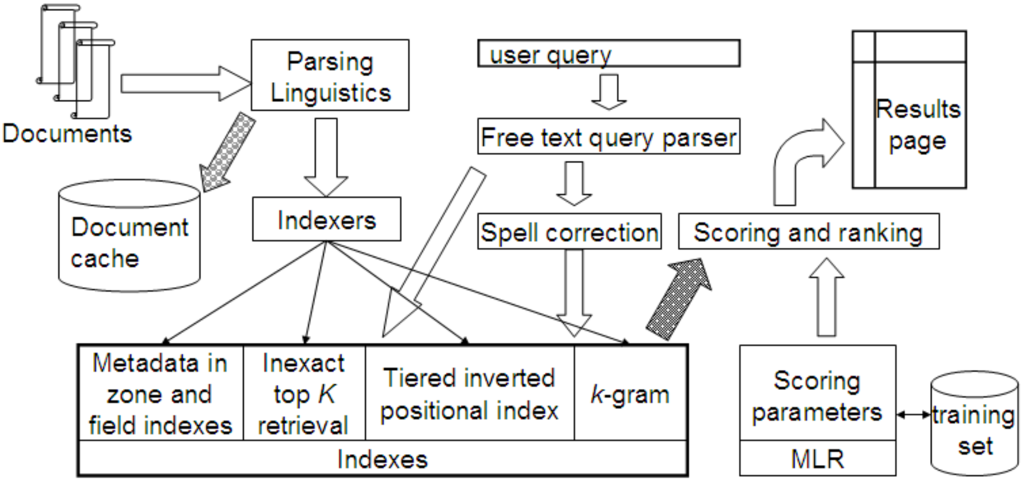
Hamid Beigy

Sharif university of technology

March 3, 2025

Table of contents

Introduction


Documents
Parsing Linguistics
user query
Free text query parser
Results page
Document cache
Indexers
Spell correction
Scoring and ranking
Metadata in zone and field indexes
Inexact top K retrieval
Tiered inverted positional index
k-gram
Indexes
Scoring parameters
MLR
training set

Introduction (framework for evaluating IR system)

Framework for the evaluation of an IR system:

1. *Test collection* consisting of
   - a document collection,
   - a test suite of information needs,
   - a set of relevance judgments for each *doc-query* pair
2. *Gold-standard* judgment of relevance.

   The classification of a document either as relevant or as irrelevant wrt an information need
3. The test collection must cover at least *50 information needs*
4. The Development collection for parameter tuning, if you need.

Standard test collections

Standard test collection

1. **Cranfield collection**: 1398 abstracts of journal articles about aerodynamics, gathered in UK in the 1950s, plus 255 queries and exhaustive relevance judgments
2. **TREC** (Text REtrieval Conference): collection maintained by the US National Institute of Standards and Technology since 1992
   - **TREC Ad Hoc Track**: test collection used for 8 evaluation campaigns led from 1992 to 1999, contains 1.89 million documents and relevance judgments for 450 topics
   - **TREC 6-8**: test collection providing 150 information needs over 528000 newswires
     - current state-of-the-art test collection
     - note that the relevance judgments

   are not exhaustive

Standard test collection

1. **GOV2**: collection also maintained by the NIST, containing 25 millions of web–pages (larger than other test collections, but smaller than current collection supported by WWW search engines)
2. **NTCIR** (Nii Test Collection for IR systems): various test collections focusing on East Asian languages, mainly used for cross-language IR
3. **CLEF** (Cross Language Evaluation Forum): collection focussing on European languages

   `http://www.clef-campaign.org`
4. **REUTERS**: Reuters 21578 and REUTERS RCV1 containing respectively 21 578 newswire articles and 806 791 documents, mainly used for text classification

Evaluation for unranked retrieval

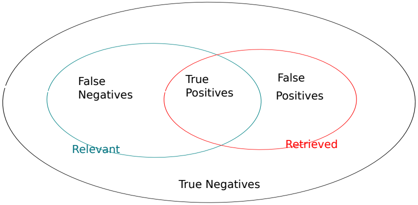
1. Two basic effectiveness measures: *precision* and *recall*

$$Pr = \frac{\#relevant\ retrieved}{\#retrieved}$$

$$Re = \frac{\#relevant\ retrieved}{\#relevant}$$

2. In other terms:

	Relevant	Not relevant
Retrieved	true positive (tp)	false positive (fp)
Not retrieved	false negative (fn)	true negative (tn)



$$Pr = \frac{tp}{tp + fp}$$

$$Re = \frac{tp}{tp + fn}$$

Evaluation for unranked retrieval: basics (continued)

1. *Accuracy*: proportion of the classification relevant/not relevant that is correct

$$accuracy = \frac{tp + tn}{tp + fp + tn + fn}$$

   Problem: 99.9% of the collection is usually not relevant to a given query (potential high rate of false positives)

2. Recall and precision are inter-dependent measures:
   - precision usually decreases while the number of retrieved documents increases
   - recall increases while the number of retrieved documents increases

Evaluation for unranked retrieval: F-measure

1. Measure relating precision and recall:

$$F = \frac{1}{\alpha \times \frac{1}{Pr} + (1-\alpha) \times \frac{1}{Re}} = \frac{(\beta^2+1) Pr \times Re}{\beta^2 Pr + Re}, \beta = \frac{1-\alpha}{\alpha}$$

2. Most frequently used: balanced F_1 with $\beta = 1$ (or $\alpha = 0.5$):

$$F_1 = \frac{2 \times Pr \times Re}{Pr + Re}$$

Example (Evaluation for unranked retrieval)

	Relevant	Not relevant	
Retrieved	20	40	60
Not retrieved	60	1,000,000	1,000,060
	80	1,000,040	1,000,120

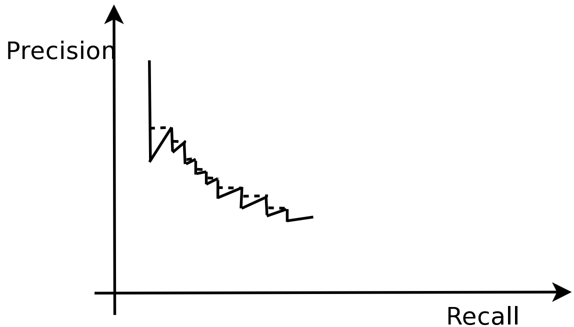
$$Pr = \frac{tp}{tp + fp} = \frac{20}{20 + 40} = \frac{1}{3}$$

$$Re = \frac{tp}{tp + fn} = \frac{20}{20 + 60} = \frac{1}{4}$$

$$F_1 = \frac{2 \times \frac{1}{3} \times \frac{1}{4}}{\frac{1}{3} + \frac{1}{4}} = \frac{2}{7}$$

Evaluation for ranked retrieval

1. Precision, recall and F-measure are set-based measures (order of documents is not important)
2. Consider the first k retrieved documents and compute the precision and recall values.
3. Plot the relation between precision and recall for each value of k
   - If the $(k+1)^{st}$ is not relevant then recall is the same, but precision decreases
   - If the $(k+1)^{st}$ is relevant then recall and precision increase
4. Precision-recall curve:

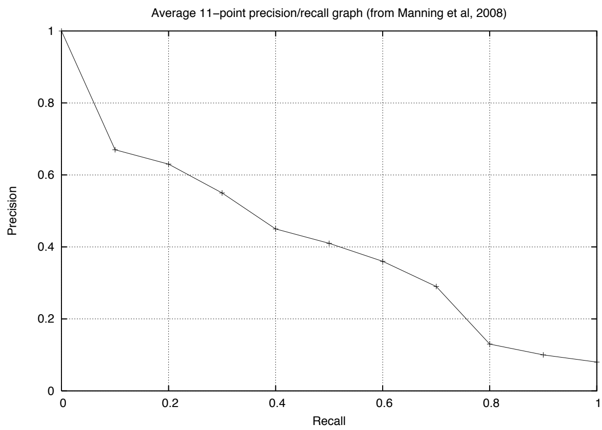


5. For removing jiggles, interpolation of the precision (smoothing):

$$P_{inter}(r) = max_{r' \geq r} P(r')$$

Ranked retrieval: efficiency measures

1. **11-point interpolated average precision**: For each information need, the value P_{inter} is measured for the 11 recall values $0.0, 0.1, 0.2, \ldots 1.0$.
2. Then, the arithmetic mean of P_{inter} for a given recall value over the information needs is computed.



1. **Precision at k**: For www search engines, we are interested in the proportion of good results among the k first answers (say the first 3 pages)

   Pros : does not need an estimate of the size of the set of relevant documents

   Cons : unstable measure, does not average well because the number of relevant documents for a query has a strong influence on precision at k.

Example (Precision at k)

Rank n	Doc
1	d_{12}
2	d_{123}
3	d_{4}
4	d_{57}
5	d_{157}
6	d_{222}
7	d_{24}
8	d_{26}
9	d_{77}
10	d_{90}

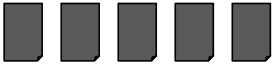
- Blue documents are relevant.
- P@n: P@3=0.33, P@5=0.2, P@8=0.25
- R@n: R@3=0.33, R@5=0.33, R@8=0.66

1. Consider rank position of each relevant doc: $k_1, k_2, \dots, k_r$.
2. Compute P@k for each $k_1, k_2, \dots, k_r$.
3. Average precision(AP) is the average of P@K.

Doc Id	Relevancy
1	r
2	n
3	r
4	n
5	r

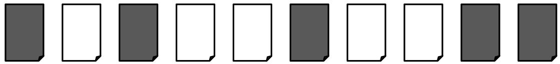
$$AP = \frac{1}{3}\left(\frac{1}{1} + \frac{2}{3} + \frac{3}{5}\right) \approx 0.76$$

4. MAP is Average Precision across multiple queries/rankings.

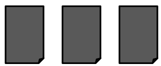


= relevant documents for query 1

Ranking #1

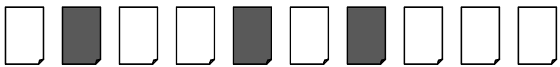


Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5



= relevant documents for query 2

Ranking #2



Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3

$$\textit{average precision query 1} = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$
$$\textit{average precision query 2} = (0.5 + 0.4 + 0.43)/3 = 0.44$$
$$\textit{mean average precision} = (0.62 + 0.44)/2 = 0.53$$

Example (MAP)

Query 1		
Rank		$P(doc_i)$
1	X	1.00
2		
3	X	0.67
4		
5		
6	X	0.50
7		
8		
9		
10	X	0.40
11		
12		
13		
14		
15		
16		
17		
18		
19		
20	X	0.25
AVG:		0.564

Query 2		
Rank		$P(doc_i)$
1	X	1.00
2		
3	X	0.67
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15	X	0.2
AVG:		0.623

$$MAP = \frac{0.564 + 0.623}{2} = 0.594$$

1. Assumptions:
   - Highly relevant documents are more useful than marginally relevant document (more likely to be examined).
   - The lower the ranked position of a relevant document, the less useful it is for the user (less likely to be examined).
2. Uses graded relevance as a measure of usefulness (gain) from examining a document.
3. Gain is accumulated starting at the top of the ranking and may be reduced, or discounted, at lower ranks
4. DCG is the total gain accumulated at a particular rank p:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log i}$$

$$DCG_p = \frac{2^{rel_{i-1}}}{log(1+i)}$$

5. Normalized Cumulative Gain (NDCG) at rank n: Normalize DCG at rank n by the DCG value at rank n of the ideal ranking.

Ranked retrieval: Normalized Discounted Cumulative Gain (NDCG)

1. Assumptions:
   - Highly relevant documents are more useful than marginally relevant document (more likely to be examined).
   - The lower the ranked position of a relevant document, the less useful it is for the user (less likely to be examined).
2. Uses graded relevance as a measure of usefulness (gain) from examining a document.
3. Gain is accumulated starting at the top of the ranking and may be reduced, or discounted, at lower ranks
4. DCG is the total gain accumulated at a particular rank p:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log i}$$

$$DCG_p = \frac{2^{rel_{i-1}}}{log(1+i)}$$

5. Normalized Cumulative Gain (NDCG) at rank n: Normalize DCG at rank n by the DCG value at rank n of the ideal ranking.

4 documents: d_1, d_2, d_3, d_4

i	Ground Truth		Ranking Function$_1$		Ranking Function$_2$	
	Document Order	r_i	Document Order	r_i	Document Order	r_i
1	d4	2	d3	2	d3	2
2	d3	2	d4	2	d2	1
3	d2	1	d2	1	d4	2
4	d1	0	d1	0	d1	0
	$NDCG_{GT}$=1.00		$NDCG_{RF1}$=1.00		$NDCG_{RF2}$=0.9203	

$$DCG_{GT} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF1} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF2} = 2 + \left(\frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.2619$$

$$MaxDCG = DCG_{GT} = 4.6309$$

Assessing relevance

1. How good is an IR system at satisfying an information need ?
2. Needs an agreement between judges $\rightarrow$ computable via the **kappa** statistic:

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

- $P(A)$ is the proportion of agreements within the judgments
- $P(E)$ is the probability that two judges agreed by chance

Example

		Judge 2		
		Yes	No	Total
Judge 1	Yes	300	20	320
	No	10	70	80
	Total	310	90	400

$$P(A) = \frac{300 + 70}{400} = \frac{370}{400} = 0.925 \qquad P(rel) = \frac{320 + 310}{400 + 400} = 0.7878$$

$$P(notrel) = \frac{80 + 90}{400 + 400} = 0.2125$$

$$P(E) = P(rel)^2 + P(notrel)^2 = (0.2125)^2 + (0.7878)^2 = 0.665$$

$$kappa = \frac{P(A) - P(E)}{1 - P(E)} = \frac{0.925 - 0.665}{1 - 0.665} = 0.776$$

1. Interpretation of the kappa statistic k:
   - $k \geq 0.8$: good agreement
   - $0.67 \leq k < 0.8$: fair agreement
   - $k < 0.67$: bad agreement
2. Note that the kappa statistic can be negative if the agreements between judgments are worse than random
3. In case of large variations between judgments, one can choose an assessor as a gold-standard

System quality and user utility

System quality and user utility

1. Ultimate interest: how satisfied is the user with the results the system gives for each of its information needs ?
2. Evaluation criteria for an IR system:
   - fast indexing
   - fast searching
   - expressivity of the query language
   - size of the collection supported
   - user interface (clearness of the input form and of the output list, *e.g.* snippets, etc)

System quality and user utility (continued)

1. Quantifying user happiness ?
   - For www search engines: do the users find the information they are looking for? can be quantified by evaluating the proportion of users getting back to the engine.
   - For intranet search engines: this efficiency can be evaluated by the time spent searching for a given piece of information.
   - General case: user studies evaluating the adequacy of the search engine with the expected usage (eCommerce, etc).

References

Reading

1. Chapters 8 of Information Retrieval Book[1]
2. Section 7.4 of Search Engines - Information Retrieval in Practice Book[2]

[1] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.
[2] W. Bruce Croft, Donald Metzler, and Trevor Strohman (2009). *Search Engines - Information Retrieval in Practice*. Pearson Education.

References

Croft, W. Bruce, Donald Metzler, and Trevor Strohman (2009). *Search Engines - Information Retrieval in Practice*. Pearson Education.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

Questions?