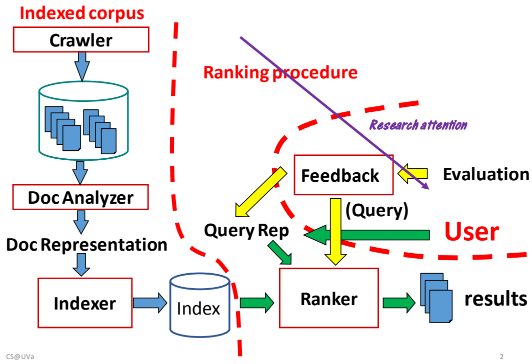
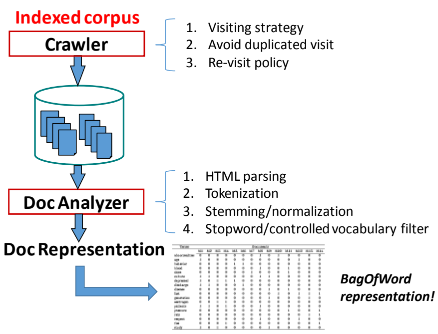
TRUY VẤN THÔNG TIN
ĐA PHƯƠNG TIỆN
INFORMATION RETRIEVAL

Evaluation IR





Nội dung

1. Tầm quan trọng của Evaluation?
2. Các tiêu chí đánh giá.
3. Một số độ đo tương ứng với bài toán.

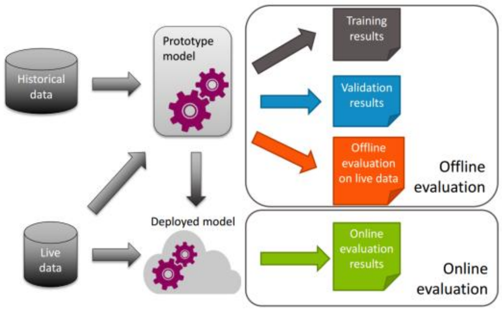
Tại sao phải đánh giá ?

1. Biết được **khi nào huấn luyện mô hình thành công** ?
2. Biết được **mức độ thành công** của mô hình
3. Biết được **thời điểm dừng quá trình huấn luyện**
4. Biết được **khi nào cần cập nhật mô hình** ?

Một số câu hỏi căn bản khi evaluation

1. Đánh giá **khi nào** ?
2. **Các tiêu chí** đánh giá là gì ?
3. Dữ liệu – **Phương pháp đánh giá** ?
4. **Độ đo nào** được sử dụng ?

When to evaluation



2. Các tiêu chí đánh giá

1. Tính chính xác (Accuracy)
2. Tính hiệu quả (Efficiency)
3. Khả năng xử lý nhiễu (Robustness).
4. Khả năng mở rộng (Scalability).
5. Khả năng diễn giải(Interpretability)
6. Mức độ phức tap (complexity)

8

2. 1 Accuracy – chính xác

➔ Tùy vào **bài toán, dữ liệu** sẽ có độ đo tương ứng.

9

2.2 Efficiency – hiệu quả

➔ Chi phí về **thời gian và tài nguyên** (bộ nhớ) cần thiết cho việc huấn luyện và kiểm thử hệ thống.

10

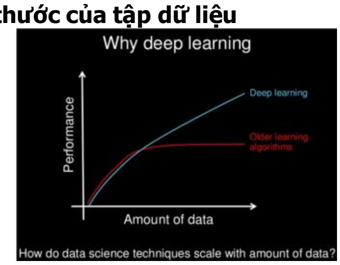
2.3 Robustnesss – xử lý nhiễu

➔ Khả năng xử lý của hệ thống đối với các ví dụ **nhiễu (lỗi)** hoặc **thiếu giá trị**.

11

2.4 Scalability – mở rộng

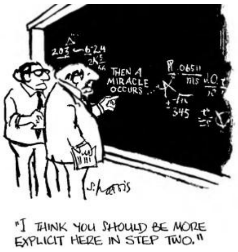
➔ **Hiệu năng** của hê thống (ví dụ: tốc độ học, độ chính xác) **thay đổi** như thế nào đối với **kích thước của tập dữ liệu**



12

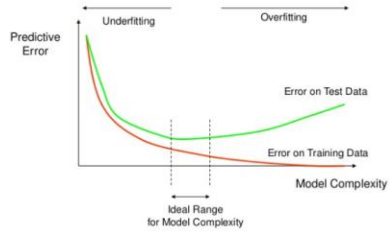
2.5 Interpretability – diễn giải

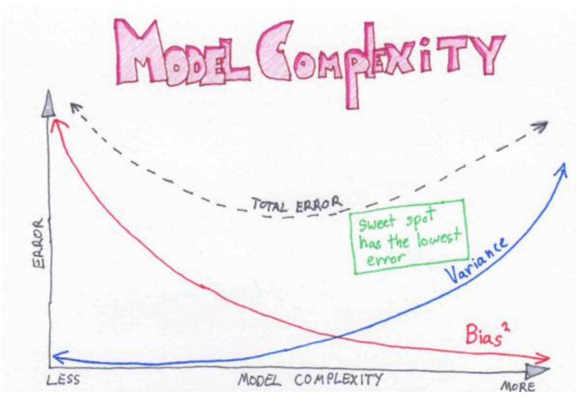
➔ **Mức độ dễ hiểu** (đối với người sử dung) của các **kết quả** và **hoạt động** của hệ thống.



2.6 Complexity – mức độ phức tạp

➔ **Mức độ phức tạp** của hệ thống (hàm hyperthesis mục tiêu) học được.





3. Một số độ đo

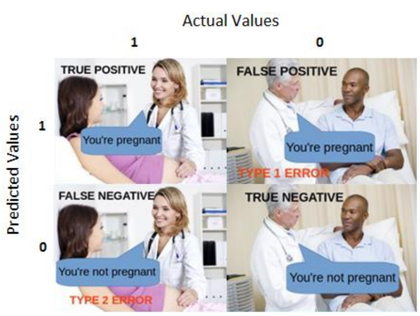
1. Accuracy/ Error
2. Precision/Recall
3. F-Score
4. AP/MAP

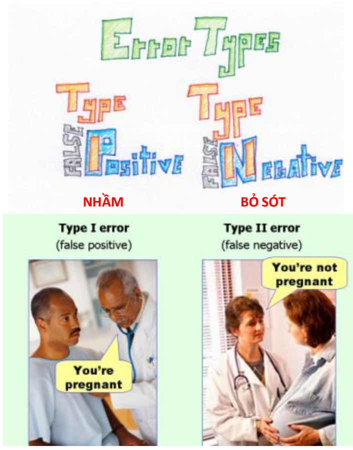
Confusion matrix (ma trận nhầm lẫn)

- ***TP****$_i$* (true positive): Số lượng các ví dụ thuộc lớp c_i được phân loại chính xác vào lớp c_i
- ***FP****$_i$* (false positive): Số lượng các ví dụ không thuộc lớp c_i bị phân loại nhầm vào lớp c_i
- ***TN****$_i$* (true negative): Số lượng các ví dụ không thuộc lớp c_i được phân loại (chính xác)
- ***FN****$_i$* (false negative): Số lượng các ví dụ thuộc lớp c_i- bị phân loại nhầm (vào các lớp khác c_i)

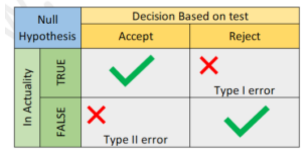
Lớp c_i		Được phân lớp **bởi hệ thống** Thuộc	Được phân lớp **bởi hệ thống** Ko thuộc
Phân lớp **thực sự (đúng)**	Thuộc	**TP$_i$**	**FN$_i$**
	Ko thuộc	**FP$_i$**	**TN$_i$**

Confusion matrix (ma trận nhầm lẫn)





Error Types

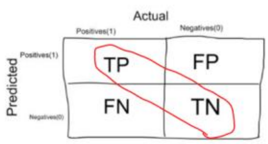


Type 1: **Loại bỏ** ví dụ mà đúng ra **không nên loại bỏ**

Type 2: **Chấp nhận** ví dụ mà đúng ra **không nên chấp nhận**

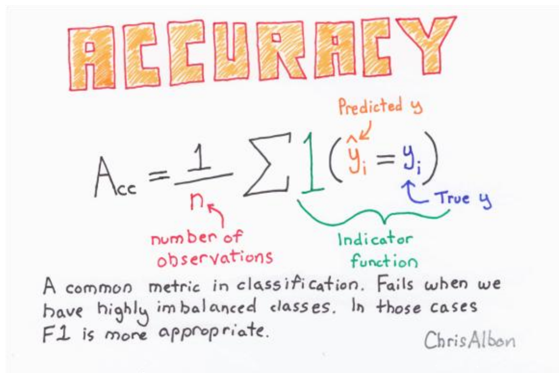
3. 1 Accuracy – độ chính xác

➔ Mức độ dự đoán (phân lớp) **chính xác** của hệ thống (đã được huấn luyện) đối với ví dụ kiểm chứng (test data).



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

Error = 1 - accuracy



3. 1 Accuracy – độ chính xác

- Là độ đo tính toán đơn giản nhất.
- Phù hợp cho các bài toán bộ dữ liệu cân bằng trong đó tỉ lệ FP (*nhầm*) và FN (*bỏ sót*) cân bằng nhau.

Hạn chế:

- Chỉ thể hiện độ chính xác không thể hiện loại lỗi trong mô hình.

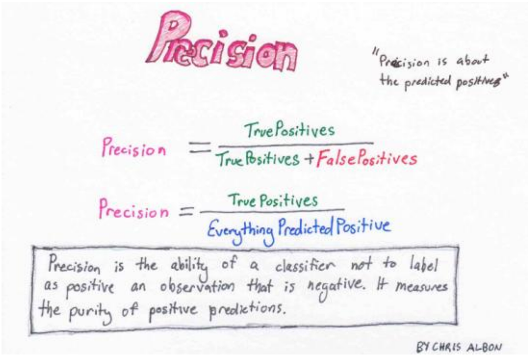
3. 2 Precision/Recall

Thực tế (Actual)

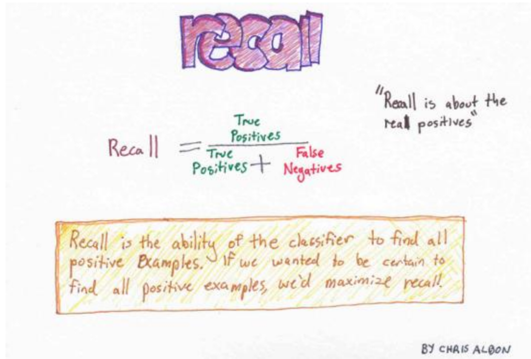
(Predicted)		1	0
	1	True Positive	False Positive
	0	False Negative	True Negative

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} = \frac{\textit{True Positive}}{\textit{Total Predicted Positive}}$$

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} = \frac{\textit{True Positive}}{\textit{Total Actual Positive}}$$



Precision được gọi là **Positive predictive value** (PPV)



Recall cũng được gọi là True Positive Rate hay Sensitivity (**độ nhạy**) – **độ phủ**

3. 2 Precision/Recall

- **Precision** đối với lớp c_i
  - → Tổng số các ví dụ thuộc lớp c_i <u>được phân loại chính xác</u> chia cho tổng số các ví dụ được phân loại vào lớp c_i

$$Precision(c_i) = \frac{TP_i}{TP_i + FP_i}$$

- **Recall** đối với lớp c_i
  - → Tổng số các ví dụ thuộc lớp c_i <u>được phân loại chính xác</u> chia cho tổng số các ví dụ thuộc lớp c_i

$$Recall(c_i) = \frac{TP_i}{TP_i + FN_i}$$

3. 2 Precision/Recall

- Làm thế nào để tính toán được giá trị Precision và Recall (một cách tổng thể) cho toàn bộ các lớp $C=\{c_i\}$?
- Trung bình vi mô (Micro-averaging)

$$Precision = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i + FP_i)} \qquad Recall = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i + FN_i)}$$

- Trung bình vĩ mô (Macro-averaging)

$$Precision = \frac{\sum_{i=1}^{|C|} Precision(c_i)}{|C|} \qquad Recall = \frac{\sum_{i=1}^{|C|} Recall(c_i)}{|C|}$$

3. 2 Precision/Recall

		Actual	
		Spam	Not Spam
Predict	Spam	8	32
	Not Spam	2	8

- Prec = 8/(8+32) = 20%
- Rec = 8/10 = 80%

➔Tỷ lệ xác suất bộ lọc chính xác khi **xác định 1 mail là thư rác** là 20%.

➔ Tỷ lệ xác suất **một thư rác bị bộ lọc phát hiện** là 80%.

3. 2 Precision/Recall

- Một mô hình tốt mong muốn khi Precision và Recall **đều cao**.
- Chọn Precision hay Recall tùy thuộc vào bài toán.

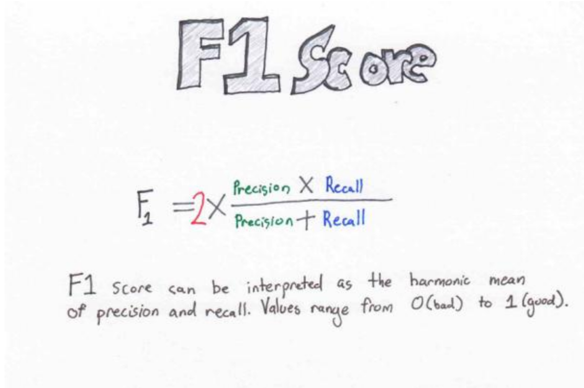
Hạn chế:

- Precision và Recall **thường mất cân bằng nhau**.

3. 3 F- Score

$$F_\beta = (1+\beta^2)\frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

- Khi β>1, recall được coi trọng hơn precision
- Khi β<1, precision được coi trọng hơn.
- Khi β=1, precision và recall coi trọng như nhau.
- β thường được sử dụng là β=2 và β=0.5

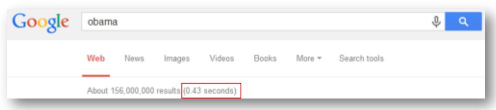


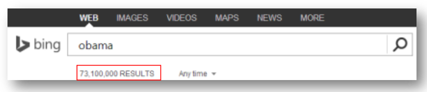
3. 3 F1 -Score

- F là một **trung bình điều hòa (harmonic mean)** của các tiêu chí Precision à Recall. Nó có xu hướng **lấy giá trị gần với giá trị nào nhỏ hơn giữa 2 tiêu chí** này.
- F1 **có giá trị lớn nếu cả 2 giá trị Precision và Recall đều lớn → F1 càng cao độ phân lớp càng tốt.**

Which search engine do you prefer: Bing or Google?

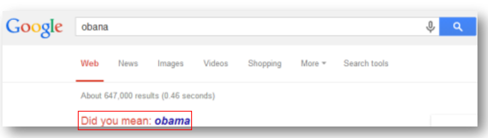
- Tiêu chuẩn đánh giá là gì ?
  - How **fast does it response** to your query?





Which search engine do you prefer: Bing or Google?

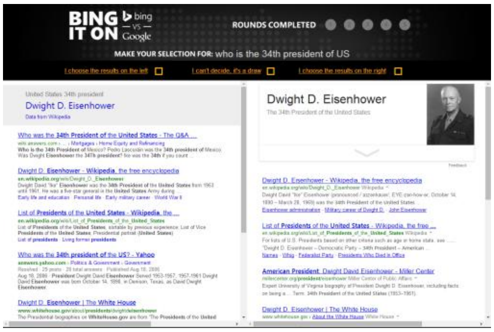
- Tiêu chuẩn đánh giá là gì ?
  - Can it correct my spelling errors?



Retrieval evaluation

- Mục tiêu của bất cứ hệ thống IR system
  - Satisfying users' information need
- Tiêu chí đo lường:
  - *"how well a system meets the information needs of its users." – wiki*
  - → *Tiêu chí này khá mô hồ và khó đo đếm*

Bing v.s. Google?



Quantify the IR quality measure

- Information need
  - *"an individual or group's desire to locate and obtain information to satisfy a conscious or unconscious need" – wiki*
  - Reflected by user query
  - Categorization of information need
    - Navigational
    - Informational
    - Transactional

Quantify the IR quality measure

- Satisfaction
  - *"the opinion of the user about a specific computer application, which they use" – wiki*
  - Reflected by
    - Increased result clicks
    - Repeated/increased visits
    - Result relevance

Classical IR evaluation

- Cranfield experiments
  - Pioneer work and foundation in IR evaluation
  - Basic hypothesis
    - Retrieved documents' relevance is a good proxy of a system's utility in satisfying users' information need
  - Procedure
    - 1,398 abstracts of aerodynamics journal articles
    - 225 queries
    - Exhaustive relevance judgments of all (query, document) pairs
    - Compare different indexing system over such collection

Classical IR evaluation

- Three key elements for IR evaluation
  1. A document collection
  2. A test suite of information needs, expressible as queries
  3. A set of relevance judgments, e.g., binary assessment of either *relevant* or *nonrelevant* for each query-document pair

Search relevance

- Users' information needs are translated into queries
- Relevance is judged with respect to the information need, **not** the query
  - E.g., Information need: "When should I renew my Virginia driver's license?"
    Query: "Virginia driver's license renewal"
    Judgment: whether a document contains the right answer, e.g., every 8 years; rather than if it literally contains those four words

Text REtrieval Conference (TREC)

- Large-scale evaluation of text retrieval methodologies
  - Since 1992, hosted by NIST
  - Standard benchmark for IR studies
  - A wide variety of evaluation collections
    - Web track
    - Question answering track
    - Cross-language track
    - Microblog track
    - And more...

Public benchmarks

TABLE 4.3 Common Test Corpora

Collection	*NDocs*	*NQrys*	*Size (MB)*	*Term/Doc*	*Q-D RelAss*
ADI	82	35			
AIT	2109	14	2	400	>10,000
CACM	3204	64	2	24.5	
CISI	1460	112	2	46.5	
Cranfield	1400	225	2	53.1	
LISA	5872	35	3		
Medline	1033	30	1		
NPL	11,429	93	3		
OSHMED	34,8566	106	400	250	16,140
Reuters	21,578	672	28	131	
TREC	740,000	200	2000	89-3543	» 100,000

Table from Manning Stanford CS276, Lecture 8

Evaluation metric

- To answer the questions
  - Is Google better than Bing?
  - Which smoothing method is most effective?
  - Is BM25 better than language models?
  - Shall we perform stemming or stopword removal?
- We need a quantifiable metric, by which we can compare different IR systems
  - As unranked retrieval sets
  - As ranked retrieval results

Recap: retrieval evaluation

- Aforementioned evaluation criteria are all good, but not essential
  - Goal of any IR system
    - Satisfying users' information need
  - Core quality measure criterion
    - *"how well a system meets the information needs of its users." – wiki*
    - Unfortunately vague and hard to execute

Recap: classical IR evaluation

- Cranfield experiments
  - Pioneer work and foundation in IR evaluation
  - Basic hypothesis
    - Retrieved documents' relevance is a good proxy of a system's utility in satisfying users' information need
  - Procedure
    - 1,398 abstracts of aerodynamics journal articles
    - 225 queries
    - Exhaustive relevance judgments of all (query, document) pairs
    - Compare different indexing system over such collection

Recap: classical IR evaluation

- Three key elements for IR evaluation
  1. A document collection
  2. A test suite of information needs, expressible as queries
  3. A set of relevance judgments, e.g., binary assessment of either *relevant* or *nonrelevant* for each query-document pair

Recap: evaluation of unranked retrieval sets

- In a Boolean retrieval system
  - Precision: fraction of retrieved documents that are relevant, i.e., p(relevant|retrieved)
  - Recall: fraction of relevant documents that are retrieved, i.e., p(retrieved|relevant)

	relevant	nonrelevant
retrieved	true positive (TP)	false positive (FP)
not retrieved	false negative (FN)	true negative (TN)

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

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

Evaluation of unranked retrieval sets

- Precision and recall trade off against each other
  - Precision decreases as the number of retrieved documents increases (unless in perfect ranking), while recall keeps increasing
  - These two metrics emphasize different perspectives of an IR system
    - Precision: prefers systems retrieving fewer documents, but highly relevant
    - Recall: prefers systems retrieving more documents

Evaluation of unranked retrieval sets

- Summarizing precision and recall to a single value
  - In order to compare different systems
  - F-measure: weighted harmonic mean of precision and recall, α balances the trade-off

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} \quad \left(F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}\right)$$

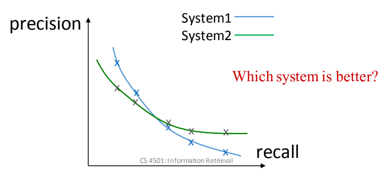
Equal weight between precision and recall

  - Why harmonic mean?
    - System1: P:0.53, R:0.36
    - System2: P:0.01, R:0.99

H	A
0.429	0.445
0.019	0.500

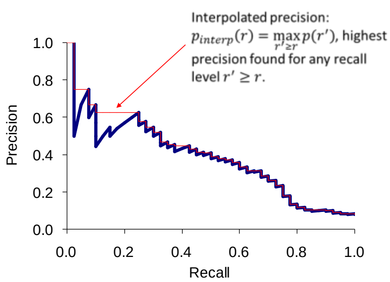
Evaluation of ranked retrieval results

- Ranked results are the core feature of an IR system
  - Precision, recall and F-measure are set-based measures, that cannot assess the ranking quality
  - Solution: evaluate precision at every recall point



Precision-Recall curve

- A sawtooth shape curve

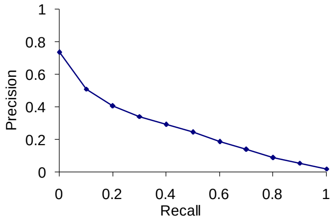


Evaluation of ranked retrieval results

- Summarize the ranking performance with a single number
  - Binary relevance
    - Eleven-point interpolated average precision
    - Precision@K (P@K)
    - Mean Average Precision (MAP)
    - Mean Reciprocal Rank (MRR)
  - Multiple grades of relevance
    - Normalized Discounted Cumulative Gain (NDCG)

Eleven-point interpolated average precision

- At the 11 recall levels [0,0.1,0.2,…,1.0], compute arithmetic mean of interpolated precision over all the queries



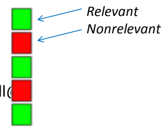
Precision@K

- Set a ranking position threshold K
- Ignores all documents ranked lower than K
- Compute precision in these top K retrieved documents
  - E.g.,:

P@3 of 2/3

P@4 of 2/4

P@5 of 3/5



- In a similar fashion we have Recall@K

Mean Average Precision

- Consider rank position of each <u>relevant</u> doc
  - E.g.,K_1, K_2, … K_R
- Compute P@K for each K_1, K_2, … K_R
- Average precision = average of those P@K
  - E.g.,

$$AvgPrec = \left(\frac{1}{1}+\frac{2}{3}+\frac{3}{5}\right)/3$$

- MAP is mean of Average Precision across multiple queries/rankings

AvgPrec is about one query

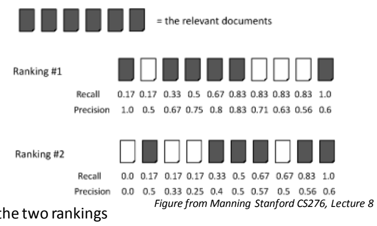


Figure from Manning Stanford CS276, Lecture 8

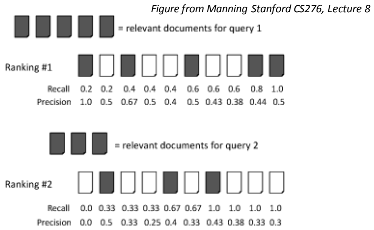
AvgPrec of the two rankings

$$\text{Ranking } \#1\text{: } (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$$

$$\text{Ranking } \#2\text{: } (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$$

MAP is about a system

Figure from Manning Stanford CS276, Lecture 8



Query 1, AvgPrec=(1.0+0.67+0.5+0.44+0.5)/5=0.62

Query 2, AvgPrec=(0.5+0.4+0.43)/3=0.44

MAP = (0.62+0.44)/2=0.53

MAP metric

- If a relevant document never gets retrieved, we assume the precision corresponding to that relevant document to be zero
- MAP is macro-averaging: each query counts equally
- MAP assumes users are interested in finding many relevant documents for each query
- MAP requires many relevance judgments in text collection

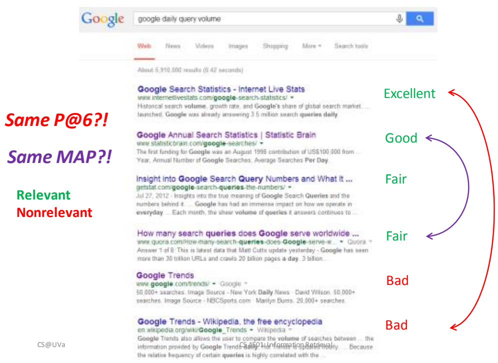
Mean Reciprocal Rank

- Measure the effectiveness of the ranked results
  - Suppose users are only looking for one relevant document
    - looking for a fact
    - known-item search
    - navigational queries
    - query auto completion
- Search duration ~ Rank of the answer
  - measures a user's effort

Mean Reciprocal Rank

- Consider the rank position, K, of the first relevant document
- Reciprocal Rank = $\frac{1}{K}$
- MRR is the mean RR across multiple queries

Beyond binary relevance



Beyond binary relevance

- The level of documents' relevance quality with respect to a given query varies
  - Highly relevant documents are more useful than marginally relevant documents
  - The lower the ranked position of a relevant document is, the less useful it is for the user, since it is less likely to be examined
  - ***Discounted Cumulative Gain***

Discounted Cumulative Gain

- Uses graded relevance as a measure of usefulness, or gain, from examining a document
- Gain is accumulated starting at the top of the ranking and discounted at lower ranks
- Typical discount is 1/log (rank)
  - With base 2, the discount at rank 4 is 1/2, and at rank 8 it is 1/3

Discounted Cumulative Gain

- DCG is the total gain accumulated at a particular rank position p:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

Relevance label at position *i*

- Alternative formulation

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(1+i)}$$

  - Standard metric in some web search companies
  - Emphasize on retrieving highly relevant documents

Normalized Discounted Cumulative Gain

- Normalization is useful for contrasting queries with varying numbers of relevant results
- Normalize DCG at rank n by the DCG value at rank n of the ideal ranking
  - The ideal ranking is achieved via ranking documents with their relevance labels

Recap: evaluation of unranked retrieval sets

- Summarizing precision and recall to a single value
  - In order to compare different systems
  - F-measure: weighted harmonic mean of precision and recall, α balances the trade-off

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \quad \left(F_1 = \frac{2}{\frac{1}{P}+\frac{1}{R}}\right)$$

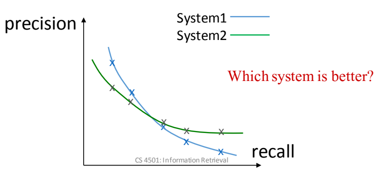
Equal weight between precision and recall

  - Why harmonic mean?
    - System1: P:0.53, R:0.36
    - System2: P:0.01, R:0.99

H	A
0.429	0.445
0.019	0.500

Recap: evaluation of ranked retrieval results

- Ranked results are the core feature of an IR system
  - Precision, recall and F-measure are set-based measures, that cannot assess the ranking quality
  - Solution: evaluate precision at every recall point



Recap: evaluation of ranked retrieval results

- Summarize the ranking performance with a single number
  - Binary relevance
    - Eleven-point interpolated average precision
    - Precision@K (P@K)
    - Mean Average Precision (MAP)
    - Mean Reciprocal Rank (MRR)
  - Multiple grades of relevance
    - Normalized Discounted Cumulative Gain (NDCG)

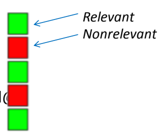
Recap: Precision@K

- Set a ranking position threshold K
- Ignores all documents ranked lower than K
- Compute precision in these top K retrieved documents
  - E.g.,:

P@3 of 2/3

P@4 of 2/4

P@5 of 3/5



- In a similar fashion we have Recall@K

Recap: Mean Average Precision

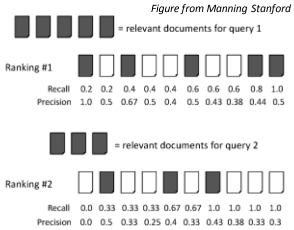
- Consider rank position of each relevant doc
  - E.g.,K_1, K_2, … K_R
- Compute P@K for each K_1, K_2, … K_R
- Average precision = average of those P@K
  - E.g.,

$$AvgPrec = \left(\frac{1}{1} + \frac{2}{3} + \frac{3}{5}\right)/3$$

- MAP is mean of Average Precision across multiple queries/rankings

Recap: MAP is about a system

Figure from Manning Stanford CS276, Lecture 8

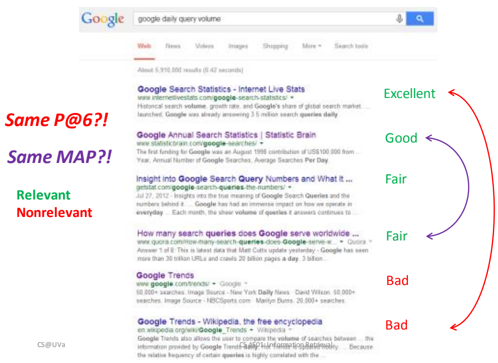


Query 1, AvgPrec=(1.0+0.67+0.5+0.44+0.5)/5=0.62
Query 2, AvgPrec=(0.5+0.4+0.43)/3=0.44
MAP = (0.62+0.44)/2=0.53

Recap: Mean Reciprocal Rank

- Measure the effectiveness of the ranked results
  - Suppose users are only looking for one relevant document
    - looking for a fact
    - known-item search
    - navigational queries
    - query auto completion
- Search duration ~ Rank of the answer
  - measures a user's effort

Recap: beyond binary relevance



Recap: Discounted Cumulative Gain

- Uses graded relevance as a measure of usefulness, or gain, from examining a document
- Gain is accumulated starting at the top of the ranking and discounted at lower ranks
- Typical discount is 1/log (rank)
  - With base 2, the discount at rank 4 is 1/2, and at rank 8 it is 1/3

Recap: Discounted Cumulative Gain

- DCG is the total gain accumulated at a particular rank position p:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

Relevance label at position *i*

- Alternative formulation

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(1+i)}$$

  - Standard metric in some web search companies
  - Emphasize on retrieving highly relevant documents

Recap: Normalized Discounted Cumulative Gain

- Normalization is useful for contrasting queries with varying numbers of relevant results
- Normalize DCG at rank n by the DCG value at rank n of the ideal ranking
  - The ideal ranking is achieved via ranking documents with their relevance labels

How about P@4, P@5, MAP and MRR?

NDCG - Example

5 documents: d_1, d_2, d_3, d_4, d_5

i	Ground Truth		Ranking Function$_1$		Ranking Function$_2$	
	Document Order	rel_i	Document Order	rel_i	Document Order	rel_i
1	d5	4	d3	2	d5	4
2	d4	3	d4	3	d3	2
3	d3	2	d2	1	d4	3
4	d2	1	d5	4	d1	0
5	d1	0	d1	0	d2	1

$$DCG_{GT} = \frac{2^4-1}{\log_2 2} + \frac{2^3-1}{\log_2 3} + \frac{2^2-1}{\log_2 4} + \frac{2^1-1}{\log_2 5} + \frac{2^0-1}{\log_2 6} = 21.35$$

$$DCG_{RF1} = \frac{2^2-1}{\log_2 2} + \frac{2^3-1}{\log_2 3} + \frac{2^1-1}{\log_2 4} + \frac{2^4-1}{\log_2 5} + \frac{2^0-1}{\log_2 6} = 14.38$$

$$DCG_{RF2} = \frac{2^4-1}{\log_2 2} + \frac{2^2-1}{\log_2 3} + \frac{2^3-1}{\log_2 4} + \frac{2^0-1}{\log_2 5} + \frac{2^1-1}{\log_2 6} = 20.78$$

What does query averaging hide?

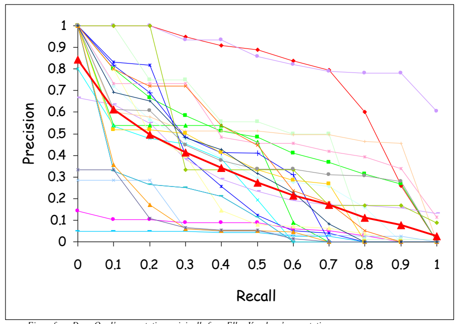


Figure from Doug Oard's presentation, originally from Ellen Voorhees' presentation

Statistical significance tests

- How confident you are that an observed difference doesn't simply result from the particular queries you chose?

Experiment 1

Query	System A	System B
1	0.20	0.40
2	0.21	0.41
3	0.22	0.42
4	0.19	0.39
5	0.17	0.37
6	0.20	0.40
7	0.21	0.41
Average	0.20	0.40

Experiment 2

Query	System A	System B
11	0.02	0.76
12	0.39	0.07
13	0.26	0.17
14	0.38	0.31
15	0.14	0.02
16	0.09	0.91
17	0.12	0.56
Average	0.20	0.40

Background knowledge

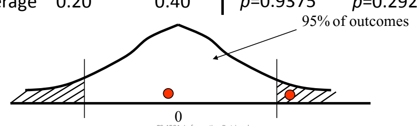
- *p*-value in statistic test is the probability of obtaining data as extreme as was observed, if the null hypothesis were true (e.g., if observation is totally random)
- If *p*-value is smaller than the chosen significance level (α), we reject the null hypothesis (e.g., observation is not random)
- We seek to reject the null hypothesis (we seek to show that the observation is a random result), and so small *p*-values are good

Tests usually used in IR evaluations

- Sign test
  - Hypothesis: the difference median is zero between samples from two continuous distributions
- Wilcoxon signed rank test
  - Hypothesis: data are paired and come from the same population
- Paired *t*-test
  - Hypothesis: difference between two responses measured on the same statistical unit has a zero mean value
- One-tail v.s. two-tail?
  - If you aren't sure, use two-tail

Statistical significance testing

Query	System A	System B	Sign Test	paired t-test
1	0.02	0.76	+	+0.74
2	0.39	0.07	-	-0.32
3	0.26	0.17	-	-0.09
4	0.38	0.31	-	-0.07
5	0.14	0.02	-	-0.12
6	0.09	0.91	+	+0.82
7	0.12	0.56	+	+0.44
Average	0.20	0.40	p=0.9375	p=0.2927



Where do we get the relevance labels?

- Human annotation
  - Domain experts, who have better understanding of retrieval tasks
    - Scenario 1: annotator lists the information needs, formalizes into queries, and judges the returned documents
    - Scenario 2: given query and associated documents, annotator judges the relevance by inferring the underlying information need

Assessor consistency

- Is inconsistency of assessors a concern?
  - Human annotators are idiosyncratic and variable
  - Relevance judgments are subjective
- Studies mostly concluded that the inconsistency didn't affect relative comparison of systems
  - Success of an IR system depends on how good it is at satisfying the needs of these idiosyncratic humans
  - Lesk & Salton (1968): assessors mostly disagree on documents at lower ranks, but measures are more affected by top-ranked documents

Measuring assessor consistency

- *kappa* statistic
  - A measure of agreement between judges
    $$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$
    - $P(A)$ is the proportion of the times judges agreed
    - $P(E)$ is the proportion of times they would be expected to agree by chance
  - $\kappa = 1$ if two judges always agree
  - $\kappa = 0$ if two judges agree by chance
  - $\kappa < 0$ if two judges always disagree

Example of *kappa* statistic

judge 2 relevance

judge 1 relevance	Yes	No	Total
Yes	300	20	320
No	10	70	80
Total	310	90	400

$$P(A) = \frac{300 + 70}{400} = 0.925$$

$$P(E) = \left(\frac{80 + 90}{400 + 400}\right)^2 + \left(\frac{320 + 310}{400 + 400}\right)^2 = 0.2125^2 + 0.7878^2 = 0.665$$

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

Prepare annotation collection

- Human annotation is expensive and time consuming
  - Cannot afford exhaustive annotation of large corpus
  - Solution: pooling
    - Relevance is assessed over a subset of the collection that is formed from the top *k* documents returned by a number of different IR systems

Does pooling work?

- Judgments cannot possibly be exhaustive?
  - Relative rankings among the systems remain the same
- What about documents beyond top *k*?
  - Relative rankings among the systems remain the same
- A lot of research work can be done here
  - Effective pool construction
  - Depth v.s. diversity

Rethink retrieval evaluation

- Goal of any IR system
  - Satisfying users' <u>information need</u>
- Core <u>quality</u> measure criterion
  - *"how well a system meets the information needs of its users." – wiki*

What we have considered

- The ability of the system to present all relevant documents
  - Recall-driven measures
- The ability of the system to withhold non-relevant documents
  - Precision-driven measures

Challenging assumptions in classical IR evaluations

- Assumption 1
  - Queries sent to an IR system would be the same as those sent to a librarian (i.e., sentence-length request), and users want to have high recall
- Assumption 2
  - Relevance = independent topical relevance
    - Documents are independently judged, and then ranked (that is how we get the ideal ranking)

What we have not considered

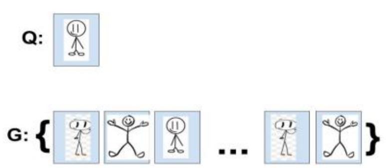
- The physical form of the output
  - User interface
- The effort, intellectual or physical, demanded of the user
  - User effort when using the system
- Bias IR research towards optimizing relevance-centric metrics

What you should know

- Core criterion for IR evaluation
- Basic components in IR evaluation
- Classical IR metrics
- Statistical test
- Annotator agreement

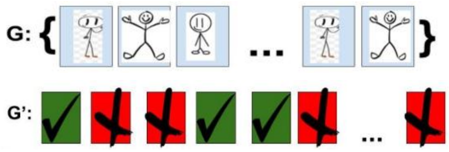
3. 4 Average Precision

Information Retrieval



- Q to be the user query
- G to be a set of labeled data in the database

3. 4 Average Precision



- Ground truth positives (GTP) – number of **True of query Q**.
- d(i,j) to be a score function to show how similar object i is to j
- G' which an ordered set of G according to score function d(,)

3. 4 Average Precision

$$\text{AP@k} = \frac{1}{\text{GTP}} \sum_{i=1}^{k} \frac{\text{TP seen}}{i}$$

- K to be the index of G'
- GTP refers to the total number of ground truth positives for the query
- TP seen refers to the number of true positives seen till k

3. 4 Average Precision

Overall AP = ⅓ (1/1 + 0/2 + 0/3 + 2/4 + ⅗ + 0 … + 0) = 0.7

Calculation of a AP for a given query, Q, with a GTP=3

3. 4 Average Precision

Overall AP = ⅓ (1/1 + 2/2 + 3/3 + 0/4 + 0/5 + 0 … + 0) = 1.0

Calculation of a pefect AP for a given query, Q, with a GTP=3

3. 4 Mean Average Precision - mAP

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveP(q)}}{Q}$$

- Q is the number of queries
- **AveP(q)** is the average precision (AP) for a given query, q

Tài liệu tham khảo

Slide được tham khảo từ:

- http://www.cs.virginia.edu/~hw5x/Course/IR2015/_site/lectures/
- https://nlp.stanford.edu/IR-book/newslides.html
- https://course.ccs.neu.edu/cs6200s14/slides.html

103