WIKIPEDIA

Evaluation measures (information retrieval)

Evaluation measures for an information retrieval system are used to assess how well the search results satisfied the user's query intent. Such metrics are often split into kinds: online metrics look at users' interactions with the search system, while offline metrics measure relevance, in other words how likely each result, or search engine results page (SERP) page as a whole, is to meet the information needs of the user.

Contents

Notation

The mathematical symbols used in the formulas below mean:

- $X \cap Y$ - Intersection - in this case, specifying the documents in *both* sets X and Y
- $|X|$ - Cardinality - in this case, the number of documents in set X

- $\int$ - [Integral](#)
- $\sum$ - [Summation](#)
- Δ - [Symmetric difference](#)

Online metrics

Online metrics are generally created from search logs. The metrics are often used to determine the success of an [A/B test](#).

Session abandonment rate

Session abandonment rate is a ratio of search sessions which do not result in a click.

Click-through rate

[Click-through rate](#) (CTR) is the ratio of users who click on a specific link to the number of total users who view a page, email, or advertisement. It is commonly used to measure the success of an [online advertising](#) campaign for a particular website as well as the effectiveness of email campaigns.[1]

Session success rate

Session success rate measures the ratio of user sessions that lead to a success. Defining "success" is often dependent on context, but for search a successful result is often measured using [dwell time](#) as a primary factor along with secondary user interaction, for instance, the user copying the result URL is considered a successful result, as is copy/pasting from the snippet.

Zero result rate

Zero result rate (*ZRR*) is the ratio of SERPs which returned with zero results. The metric either indicates a [recall](#) issue, or that the information being searched for is not in the index.

Offline metrics

Offline metrics are generally created from relevance judgment sessions where the judges score the quality of the search results. Both binary (relevant/non-relevant) and multi-level (e.g., relevance from 0 to 5) scales can be used to score each document returned in response to a query. In practice, queries may be [ill-posed](#), and there may be different shades of relevance. For instance, there is ambiguity in the query "mars": the judge does not know if the user is searching for the planet [Mars](#), the [Mars](#) chocolate bar, or the singer [Bruno Mars](#).

Precision

Precision is the fraction of the documents retrieved that are [relevant](#) to the user's information need.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

In binary classification, precision is analogous to positive predictive value. Precision takes all retrieved documents into account. It can also be evaluated considering only the topmost results returned by the system using Precision@k.

Note that the meaning and usage of "precision" in the field of information retrieval differs from the definition of accuracy and precision within other branches of science and statistics.

Recall

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

In binary classification, recall is often called sensitivity. So it can be looked at as *the probability that a relevant document is retrieved by the query*.

It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by computing the precision.

Fall-out

The proportion of non-relevant documents that are retrieved, out of all non-relevant documents available:

$$\text{fall-out} = \frac{|\{\text{non-relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{non-relevant documents}\}|}$$

In binary classification, fall-out is closely related to specificity and is equal to $(1 - \text{specificity})$. It can be looked at as *the probability that a non-relevant document is retrieved by the query*.

It is trivial to achieve fall-out of 0% by returning zero documents in response to any query.

F-score / F-measure

The weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score is:

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}$$

This is also known as the F_1 measure, because recall and precision are evenly weighted.

The general formula for non-negative real β is:

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

Two other commonly used F measures are the F_2 measure, which weights recall twice as much as precision, and the $F_{0.5}$ measure, which weights precision twice as much as recall.

The F-measure was derived by van Rijsbergen (1979) so that F_β "measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision". It is based on van Rijsbergen's effectiveness measure $E = 1 - \dfrac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}}$. Their relationship is:

$$F_\beta = 1 - E \text{ where } \alpha = \frac{1}{1 + \beta^2}$$

F-measure can be a better single metric when compared to precision and recall; both precision and recall give different information that can complement each other when combined. If one of them excels more than the other, F-measure will reflect it.

Average precision

Precision and recall are single-value metrics based on the whole list of documents returned by the system. For systems that return a ranked sequence of documents, it is desirable to also consider the order in which the returned documents are presented. By computing a precision and recall at every position in the ranked sequence of documents, one can plot a precision-recall curve, plotting precision $p(r)$ as a function of recall r. Average precision computes the average value of $p(r)$ over the interval from $r = 0$ to $r = 1$:[2]

$$\text{AveP} = \int_0^1 p(r)dr$$

That is the area under the precision-recall curve. This integral is in practice replaced with a finite sum over every position in the ranked sequence of documents:

$$\text{AveP} = \sum_{k=1}^{n} P(k)\Delta r(k)$$

where k is the rank in the sequence of retrieved documents, n is the number of retrieved documents, $P(k)$ is the precision at cut-off k in the list, and $\Delta r(k)$ is the change in recall from items $k-1$ to k.[2]

This finite sum is equivalent to:

$$\text{AveP} = \frac{\sum_{k=1}^{n} (P(k) \times \text{rel}(k))}{\text{number of relevant documents}}$$

where $\text{rel}(k)$ is an indicator function equaling 1 if the item at rank k is a relevant document, zero otherwise.[3] Note that the average is over all relevant documents and the relevant documents not retrieved get a precision score of zero.

Some authors choose to interpolate the $p(r)$ function to reduce the impact of "wiggles" in the curve.[4][5] For example, the PASCAL Visual Object Classes challenge (a benchmark for computer vision object detection) until 2010[6] computed the average precision by averaging the precision over a set of evenly spaced recall levels $\{0, 0.1, 0.2, \dots 1.0\}$:[4][5]

$$\text{AveP} = \frac{1}{11} \sum_{r \in \{0, 0.1, \dots, 1.0\}} p_{\text{interp}}(r)$$

where $p_{\text{interp}}(r)$ is an interpolated precision that takes the maximum precision over all recalls greater than r:

$$p_{\text{interp}}(r) = \max_{\tilde{r} : \tilde{r} \geq r} p(\tilde{r}).$$

An alternative is to derive an analytical $p(r)$ function by assuming a particular parametric distribution for the underlying decision values. For example, a *binormal precision-recall curve* can be obtained by assuming decision values in both classes to follow a Gaussian distribution.[7]

Precision at k

For modern (web-scale) information retrieval, recall is no longer a meaningful metric, as many queries have thousands of relevant documents, and few users will be interested in reading all of them. Precision at k documents (P@k) is still a useful metric (e.g., P@10 or "Precision at 10" corresponds to the number of relevant results among the top 10 retrieved documents), but fails to take into account the positions of the relevant documents among the top k.[8] Another shortcoming is that on a query with fewer relevant results than k, even a perfect system will have a score less than 1.[9] It is easier to score manually since only the top k results need to be examined to determine if they are relevant or not.

R-Precision

R-precision requires knowing all documents that are relevant to a query. The number of relevant documents, R, is used as the cutoff for calculation, and this varies from query to query. For example, if there are 15 documents relevant to "red" in a corpus (R=15), R-precision for "red" looks at the top 15 documents returned, counts the number that are relevant r turns that into a relevancy fraction: $r/R = r/15$.[10]

R-Precision is equal to recall at the **R**-th position.[9]

Empirically, this measure is often highly correlated to mean average precision.[9]

Mean average precision

Mean average precision for a set of queries is the mean of the average precision scores for each query.

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

where *Q* is the number of queries.

Discounted cumulative gain

DCG uses a graded relevance scale of documents from the result set to evaluate the usefulness, or gain, of a document based on its position in the result list. The premise of DCG is that highly relevant documents appearing lower in a search result list should be penalized as the graded relevance value is reduced logarithmically proportional to the position of the result.

The DCG accumulated at a particular rank position p is defined as:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

balanced accuracy (BA)

$$\text{BA} = \frac{TPR + TNR}{2}$$

F1 score

is the harmonic mean of precision and sensitivity:

$$\text{F}_1 = 2 \times \frac{\text{PPV} \times \text{TPR}}{\text{PPV} + \text{TPR}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

Matthews correlation coefficient (MCC)

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

Fowlkes–Mallows index (FM)

$$\text{FM} = \sqrt{\frac{TP}{TP + FP} \times \frac{TP}{TP + FN}} = \sqrt{PPV \times TPR}$$

informedness or bookmaker informedness (BM)

$$\text{BM} = \text{TPR} + \text{TNR} - 1$$

markedness (MK) or deltaP (Δp)

$$\text{MK} = \text{PPV} + \text{NPV} - 1$$

Sources: Fawcett (2006),[11] Piryonesi and El-Diraby (2020),[12] Powers (2011),[13] Ting (2011),[14] CAWCR,[15] D. Chicco & G. Jurman (2020, 2021),[16][17] Tharwat (2018).[18]

$$\text{DCG}_\text{p} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)}.$$

Since result set may vary in size among different queries or systems, to compare performances the normalised version of DCG uses an ideal DCG. To this end, it sorts documents of a result list by relevance, producing an ideal DCG at position p ($IDCG_p$), which normalizes the score:

$$\text{nDCG}_\text{p} = \frac{DCG_p}{IDCGp}.$$

The nDCG values for all queries can be averaged to obtain a measure of the average performance of a ranking algorithm. Note that in a perfect ranking algorithm, the DCG_p will be the same as the $IDCG_p$ producing an nDCG of 1.0. All nDCG calculations are then relative values on the interval 0.0 to 1.0 and so are cross-query comparable.

Other measures

- Mean reciprocal rank
- Spearman's rank correlation coefficient
- bpref - a summation-based measure of how many relevant documents are ranked before irrelevant documents[10]
- GMAP - geometric mean of (per-topic) average precision[10]
- Measures based on marginal relevance and document diversity - see Relevance (information retrieval) § Problems and alternatives
- Measures of both relevance and credibility (for fake news in search results)[19]

Visualization

Visualizations of information retrieval performance include:

- Graphs which chart precision on one axis and recall on the other[10]
- Histograms of average precision over various topics[10]
- Receiver operating characteristic (ROC curve)
- Confusion matrix

Non-metrics

Top queries list

Top queries is noting the most common queries over a fixed amount of time. The top queries list assists in knowing the style of queries entered by users.

Non-relevance metrics

Queries per time

Measuring how many queries are performed on the search system per (month/day/hour/minute/sec) tracks the utilization of the search system. It can be used for diagnostics to indicate an unexpected spike in queries, or simply as a baseline when comparing with other metrics, like query latency. For example, a spike in query traffic, may be used to explain a spike in query latency.

See also

- Universal IR Evaluation
- Learning to rank

References

1. American Marketing Association Dictionary. [1] (https://web.archive.org/web/20080527180926/http://www.marketingpower.com/_layouts/Dictionary.aspx?) Retrieved 2012-11-02. The Marketing Accountability Standards Board (MASB) endorses this definition as part of its ongoing Common Language in Marketing Project (http://www.commonlanguage.wikispaces.net/).
2. Zhu, Mu (2004). "Recall, Precision and Average Precision" (https://web.archive.org/web/20110504130953/http://sas.uwaterloo.ca/stats_navigation/techreports/04WorkingPapers/2004-09.pdf) (PDF). Archived from the original (http://sas.uwaterloo.ca/stats_navigation/techreports/04WorkingPapers/2004-09.pdf) (PDF) on 2011-05-04.

3. Turpin, Andrew; Scholer, Falk (2006). *User performance versus precision measures for simple search tasks* (https://archive.org/details/sigirseattle20060000inte/page/11). *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Seattle, WA, August 06–11, 2006)*. New York, NY: ACM. pp. 11–18 (http s://archive.org/details/sigirseattle20060000inte/page/11). CiteSeerX 10.1.1.533.4100 (https://cit eseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.533.4100). doi:10.1145/1148170.1148176 (ht tps://doi.org/10.1145%2F1148170.1148176). ISBN 978-1-59593-369-0. S2CID 9810253 (http s://api.semanticscholar.org/CorpusID:9810253).

4. Everingham, Mark; Van Gool, Luc; Williams, Christopher K. I.; Winn, John; Zisserman, Andrew (June 2010). "The PASCAL Visual Object Classes (VOC) Challenge" (https://web.archive.org/ web/20111120074515/http://pascallin.ecs.soton.ac.uk/challenges/VOC/pubs/everingham10.pd f) (PDF). *International Journal of Computer Vision*. **88** (2): 303–338. doi:10.1007/s11263-009-0275-4 (https://doi.org/10.1007%2Fs11263-009-0275-4). hdl:20.500.11820/88a29de3-6220-442b-ab2d-284210cf72d6 (https://hdl.handle.net/20.500.11820%2F88a29de3-6220-442b-ab2d-284210cf72d6). S2CID 4246903 (https://api.semanticscholar.org/CorpusID:4246903). Archived from the original (http://pascallin.ecs.soton.ac.uk/challenges/VOC/pubs/everingham1 0.pdf) (PDF) on 2011-11-20. Retrieved 2011-08-29.

5. Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich (2008). *Introduction to Information Retrieval* (http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retri eval-results-1.html). Cambridge University Press.

6. "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Development Kit" (http://hos t.robots.ox.ac.uk/pascal/VOC/voc2012/htmldoc/devkit_doc.html#sec:ap). *host.robots.ox.ac.uk*. Retrieved 2019-03-23.

7. K.H. Brodersen, C.S. Ong, K.E. Stephan, J.M. Buhmann (2010). The binormal assumption on precision-recall curves (http://icpr2010.org/pdfs/icpr2010_ThBCT8.28.pdf) Archived (https://we b.archive.org/web/20121208201457/http://icpr2010.org/pdfs/icpr2010_ThBCT8.28.pdf) December 8, 2012, at the Wayback Machine. *Proceedings of the 20th International Conference on Pattern Recognition*, 4263-4266.

8. Kalervo, J~irvelin (2017). "IR evaluation methods for retrieving highly relevant documents" (http s://sigir.org/wp-content/uploads/2017/06/p243.pdf) (PDF). *ACM SIGIR Forum*. **51, 2**: 243–250.

9. Christopher D. Manning; Prabhakar Raghavan & Hinrich Schütze (2009). "Chapter 8: Evaluation in information retrieval" (http://nlp.stanford.edu/IR-book/pdf/08eval.pdf) (PDF). Retrieved 2015-06-14. Part of *Introduction to Information Retrieval* [2] (http://nlp.stanford.edu/IR-book/)

10. http://trec.nist.gov/pubs/trec15/appendices/CE.MEASURES06.pdf

11. Fawcett, Tom (2006). "An Introduction to ROC Analysis" (http://people.inf.elte.hu/kiss/11dwhd m/roc.pdf) (PDF). *Pattern Recognition Letters*. **27** (8): 861–874. doi:10.1016/j.patrec.2005.10.010 (https://doi.org/10.1016%2Fj.patrec.2005.10.010).

12. Piryonesi S. Madeh; El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". *Journal of Infrastructure Systems*. **26** (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512 (https://doi.org/10.1061%2F%28 ASCE%29IS.1943-555X.0000512).

13. Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (https://www.researchgate.net/publication/22852930 7). *Journal of Machine Learning Technologies*. **2** (1): 37–63.

14. Ting, Kai Ming (2011). Sammut, Claude; Webb, Geoffrey I. (eds.). *Encyclopedia of machine learning*. Springer. doi:10.1007/978-0-387-30164-8 (https://doi.org/10.1007%2F978-0-387-301 64-8). ISBN 978-0-387-30164-8.

15. Brooks, Harold; Brown, Barb; Ebert, Beth; Ferro, Chris; Jolliffe, Ian; Koh, Tieh-Yong; Roebber, Paul; Stephenson, David (2015-01-26). "WWRP/WGNE Joint Working Group on Forecast Verification Research" (https://www.cawcr.gov.au/projects/verification/). *Collaboration for Australian Weather and Climate Research*. World Meteorological Organisation. Retrieved 2019-07-17.

16. Chicco D., Jurman G. (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). *BMC Genomics*. **21** (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7 (https://doi.org/10.1186%2Fs12864-019-6413-7). PMC 6941312 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). PMID 31898477 (https://pubmed.ncbi.nlm.nih.gov/31898477).

17. Chicco D., Toetsch N., Jurman G. (February 2021). "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7863449). *BioData Mining*. **14** (13): 1-22. doi:10.1186/s13040-021-00244-z (https://doi.org/10.1186%2Fs13040-021-00244-z). PMC 7863449 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7863449). PMID 33541410 (https://pubmed.ncbi.nlm.nih.gov/33541410).

18. Tharwat A. (August 2018). "Classification assessment methods" (https://doi.org/10.1016%2Fj.aci.2018.08.003). *Applied Computing and Informatics*. doi:10.1016/j.aci.2018.08.003 (https://doi.org/10.1016%2Fj.aci.2018.08.003).

19. C. Lioma; J. G. Simonsen; B. Larsen (2017). "Evaluation Measures for Relevance and Credibility in Ranked Lists" (http://main.cl-lab.dk/www/publications/2017/pdf/a66.pdf) (PDF). *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, 91-98.