Measurement of Relevance Effectiveness

Relevance

- "a state or quality of being to the purpose;
- a state or quality of being related to the subject or matter at hand"

[The Cambridge English Dictionary, Grandreams Limited, London, English Edition, 1990]

Measures

The *effectiveness* of an information retrieval system means how well (or bad) it performs.

Effectiveness measures are elaborated based on different categories such as:

- Relevance,
- Efficiency,
- Utility,
- User satisfaction.

Within each category, there are different specific effectiveness measures:

- **Relevance:**
  - precision,
  - recall,
  - fallout, etc.,
- **Efficiency:**
  - cost of search,
  - amount of search time, etc.,
- **Utility:**
  - worth of search results in some currency,
  - etc.,
- **User satisfaction:**
  - user's satisfaction with precision

- or intermediary's understanding of request,
- etc..

Relevance effectiveness

is the ability of a retrieval method or system

to return relevant answers.

The traditional measures are the following:

- *Precision*: the proportion of relevant documents out of those returned.
- *Recall*: the proportion of returned documents out of the relevant ones.
- *Fallout*: the proportion of returned documents out of those nonrelevant.

Let D denote a collection of documents, q a query, and

- $\Delta \neq 0$ denote the total number of relevant documents to query q,
- $\kappa \neq 0$ denote the number of retrieved documents in response to query q,
- α denote the number of retrieved and relevant documents.

It is reasonable to assume that the total number of documents to be searched, M, is greater than those retrieved,

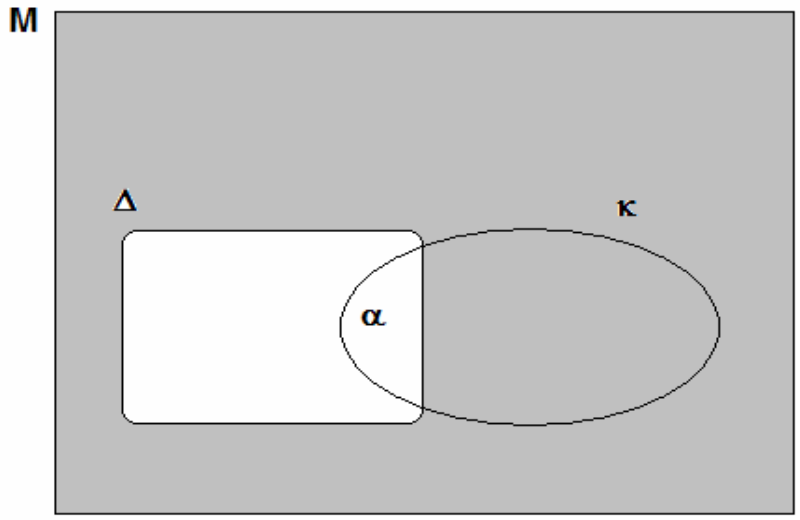
i.e., $|D| = M > \Delta$.

Then, the usual relevance effectiveness measures are defined formally as follows:

1. *Recall* ρ is defined as $\rho = \frac{\alpha}{\Delta}$

2. *Precision* π is defined as $\pi = \frac{\alpha}{\kappa}$

3. *Fallout* φ is defined as $\varphi = \frac{\kappa - \alpha}{M - \Delta}$



Visual representation of quantities which define precision, recall, fallout

Properties:

- $0 \leq \rho \leq 1,$
- $0 \leq \pi \leq 1,$
- $\rho = 0 \Leftrightarrow \pi = 0,$
- $\pi = 1 \Leftrightarrow \varphi = 0,$
- $\alpha = \kappa = \Delta \Leftrightarrow (\rho = \pi = 1 \wedge \varphi = 0).$

Other measures:

$$\pi + \rho,$$

$$\pi + \rho - 1,$$

$$\frac{\rho - \varphi}{\rho + \varphi - 2\rho\varphi}, \quad 0 \leq \varphi \leq 1,$$

$$1 - \frac{1}{\frac{1}{2}\left(\frac{1}{\pi}\right) - \frac{1}{2}\left(\frac{1}{\rho}\right)},$$

F-measure: $\dfrac{2\rho\pi}{\rho+\pi}$,

Heine measure: $1-\dfrac{1}{\dfrac{1}{\pi}+\dfrac{1}{\rho}-1}$,

Vickery measure: $1-\dfrac{1}{2\left(\dfrac{1}{\pi}\right)+2\left(\dfrac{1}{\rho}\right)-3}$,

Meadow measure: $1-\dfrac{\sqrt{(1-\pi)^2+(1-\rho)^2}}{\sqrt{2}}$.

$$R_{norm}=\frac{1}{M-\Delta}\sum_{i=1}^{M}\rho_i-\frac{\Delta+1}{2(M-\Delta)},$$ where

M: the number of documents,

R_{norm}: normalised recall (for a given query),

ρ_i: recall at the ith hit in the ranked hit list.

Precision-Recall Graph Method

The *precision-recall graph method* is being used for the measurement of retrieval effectiveness under laboratory conditions, i.e, in a controlled and repeatable manner.

In this measurement method, *test databases* (*test collections*) are used. Each test collection is manufactured by specialists, and has a fixed structure as follows:

- The documents *d* are given.
- The queries *q* are given.
- The relevance list is given, i.e., it is exactly known which document is relevant to which query.

1. For every query, retrieval should be performed (using the retrieval method whose relevance effectiveness is to be measured).
2. The hit list is compared with the relevance list (corresponding to the query under focus).

   The following recall levels are considered to be standard levels:

   0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9; 1;

   (the levels can also be given as %, for example 0.1 = 10%).
3. For every query, pairs of recall and precision are computed.
4. If the computed recall value is not standard, then it is approximated.
5. The precision values corresponding to equal recall values are averaged.

Let

R_q denote the relevant documents to the query q. Let us assume, for example, that

$$R_q = \{d_2, d_4, d_6, d_5, d_9, d_1\}, \quad \Delta = 6.$$

Let us assume that the retrieval method under measurement returns the following ranked hit list:

1. d_1—
2. d_8
3. d_6—
4. d_7
5. d_9—

where the "—" sign marks a relevant document.

- The document d_1 is relevant. This means that 1/6th of the documents of R_q have been retrieved, and so precision is 100% at the recall level 1/6.
- The fifth element of the hit list is d_9 which is also relevant. Hence, precision is 3/5 = 0.6 at the recall level 3/6 = 0.5.

When the computed recall value r is not equal to a standard level, the following interpolation method can be used to calculate the precision value $p(r_j)$ corresponding to the standard recall value r_j:

$$p(r_j) = \max_{r_{j-1}<r\leq r_j} p(r)$$

where r_j, $j = 2,\ldots,10$, denotes the j^{th} standard recall level. It is known from practice that the values $p(r_j)$ are monotonically decreasing.

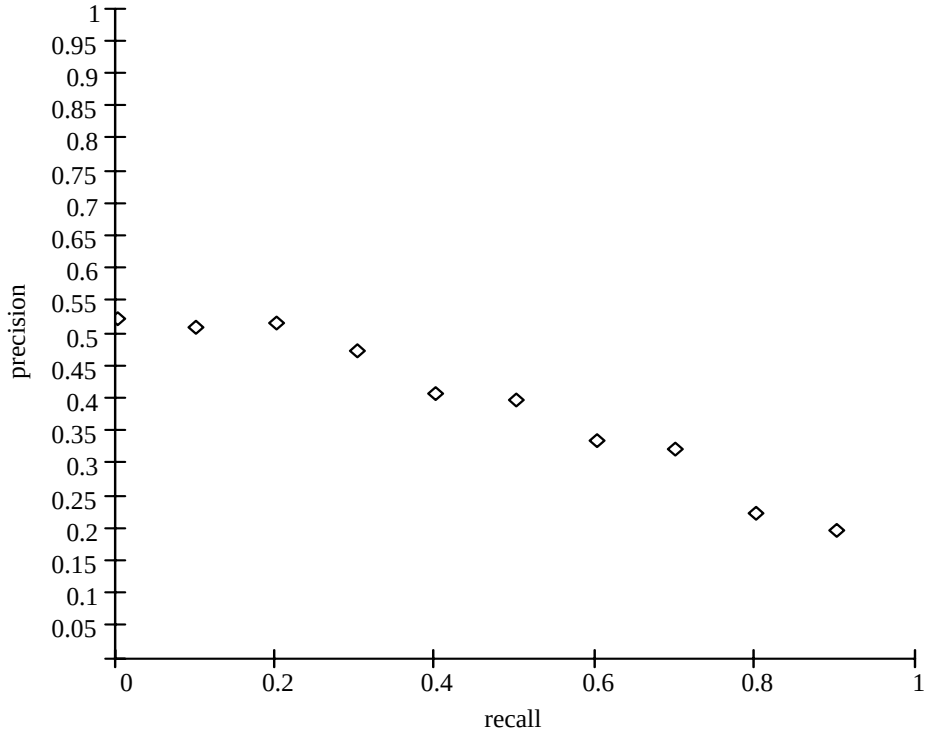
Thus, the value $p(r_1)$ is usually so determined as

$$p(r_1) \geq p(r_2).$$

For all queries q_i, the precision values $p_i(r_j)$ are averaged at all standard recall levels as follows:

$$P(r_j) = \frac{1}{n} \sum_{i=1}^{n} p_i(r_j), \quad j = 1,...,10,$$

where n denotes the number of queries used.



typical precision-recall graph

The average of the values $P(r_j)$ is called MAP (*Mean Average Precision*).

MAP can also be computed just at the recall values 0.3, 0.6, and 0.9.

Apart from MAP, the following measures can also be used:

- *P@n* (*precision* at *n*): only the first *n* elements of every hit list is considered; typical values for *n* are $n = 10, 20, 30, 100$.
- *R-prec* (*R precision*): for each query *q*, only the first Δ_q elements of the hit list is considered (i.e., $\Delta_q = R_q$).

Measurement of Search Engine Effectiveness

- The measurement of relevance effectiveness of a Web search engine is, typically (due to the characteristics of the Web), user centred
- It is an experimentally established fact that the majority of users examine, in general, the first two page of a hit list.
- Thus, the search engine should rank the most relevant pages in the first few pages.

The traditional measures cannot always be computed (for example, recall and fallout).

⇓

search engines requires other measures than the traditional ones.

When elaborating such new measures:

- one is trying to use traditional measures (for example, precision which can be calculated also for a hit list of a search engine), and
- on the other hand, takes into account different characteristics of the Web.

Several methods for the measurement of relevance effectiveness of a search engine have been elaborated thus far. They can be grouped as follows:

User-based methods. These method measure user satisfaction.

20 full precision =

$$\frac{1}{20 \times 4} \sum_{i=1}^{20} (weight\_of\_i^{th}\_hit)$$

M-L-S Method

The principles of the method are as follows:

- definition of relevance categories,
- definition of groups,
- weighting of hits.

Each hit of a hit list returned in response to a query was assigned to exactly one category.

The hit list was divided into s_i groups having c_i weights (i = 1,...,m).

The value of first n-precision was defined as the sum of the weights of relevant hits divided by the maximum sum.

The M-L-S method measures the capability of a search engine to rank relevant hits within the first 5 or 10 hits of the hit list.

The M-L-S method is as follows.

M-L-S method (first 5/10-precision)

1. Select search engine to be measured.
2. Define relevance categories.
3. Define groups.
4. Define weights.
5. Give queries q_i (i = 1,...,s).
6. Compute $P5_i$ and/or $P10_i$ for q_i (i=1,...,s).
7. The first 5/10-precision of the search engine is:

$$Pk = \frac{1}{s}\sum_{i=1}^{s} Pk_i \text{ , where } k = 5 \text{ or } k = 10.$$

The relevance categories are as follows:

- 0-category (irrelevant hit),
- 1-category (relevant hit).

When measuring **first 5-precision**, the first five hits are grouped into two groups as follows:

1. group: the first two hits (on the ground that they are on the first screen),
2. group: the following three hits.

When measuring the **first 10-precision**, the first ten hits are grouped into the following three groups:

1. group: the first two hits,
2. group: the next three hits,
3. group: the rest of five hits.

- Groups 1 and 2 are based on the assumption that, in practice, the most important hits are the first five (usually on the first screen).
- Hits within the same group get equal weights.
- The weights reflect that the user is more satisfied if the relevant hits appear on the first screen.

For the first 5-precision, the weights are:

1. For group 1: 10.
2. For group 2: 5.

For the first 10-precision, the weights are:

1. For group 1: 20.
2. For group 2: 17.
3. For group 3: 10.

Instead of 20, 17, 10, (or 10, 5 for the first 5-precision) other but proportional values may be used.

The definition of queries is a very important step.

- It is advisable to define a topic first, and the queries after that. The topic should be broad enough as the goal is to see how well the search engine performs at a general level.
- In order to avoid bias, define both general and specialised queries.
- As most users prefer unstructured queries, such queries should be defined.
- It is very important that the weights be defined prior to obtaining any hits, or else our assessments would be more subjective or biased (because, in this case, we already get to know how the search engine ‚behaves' for certain queries).

The $P5$ measure is defined as follows:

$$P5 = \frac{no\_relevant\_hits_{1.-2.hit} \times 10 + no\_relevant\_hits_{3.-5.hit} \times 5}{35 - ((5 - no\_hits_{1.-5.hit}) \times 5)},$$

where

- the numerator is the weighted sum of the relevant hits within the first five hits,
- in the denominator, 35 is the weighted sum in the best case (i.e., when the first five hits are all relevant): $(2 \times 10) + (3 \times 5) = 35$. For every missing hit out of five, 5 is subtracted.

The measure $P5$ is given for the case when multiple hits are not penalised.

If we want to penalise multiple hits, then a multiple hit is considered as many different hits as its multiplicity.

Example

1. Let us assume that in response to the query "WWW" 3 hits are returned, and that all are relevant.

   Thus, the numerator is $(2 \times 10) + (1 \times 5) = 25$.

   The first two hits belong to the first group, so their weight is 10.

   The third hit belongs to group 2, thus its weight is 5.

   The denominator is $35 - (2 \times 5) = 25$.

   So, $P5 = 25 : 25 = 1$.

2.For the query be “VLSI” five hits are returned, out of which three are relevant: 2., 3., and 4.. Thus, the numerator is $(1 \times 10) + (2 \times 5) = 20$, and so $P5 = 20 : 35 = 0.571$.

If the first three hits were relevant, then $P5 = ((2 \times 10) + (1 \times 5)) : 35 = 0.714$.

The two values obtained for $P5$ are different, which reflects the ranking difference of relevant hits.

3.For the query “Network” five hits are returned, and these are relevant, but the third and the fifth are the same (that is we have a double hit).

In this case, we have $P5 = ((2 \times 10) + (2 \times 5)) : (35 – 1 \times 5) = 1$ (without penalty);and

$P5 = ((2 \times 10) + (2 \times 5)) : 35 = 0.857$ (with penalty).

The $P10$ measure is defined in a similar manner as follows:

$$P10 = \frac{r\_hit_{1.-2.hit} \times 20 + r\_hit_{3.-5.hit} \times 17 + r\_hit_{6.-10.hit} \times 10}{141 - ((10 - no - hits_{1.-10.link}) \times 10)},$$

where

r_hit denotes the number of relevant hits in the respective group.

The penalised version is similar to what was said for $P5$.

RP Method

We know that precision is defined as follows

$$p = \frac{r}{k},$$

where

p denotes precision,

k the number of returned items,

r the relevant items out of these *k* returned.

A Web meta-search engine uses the hit lists of search engines to produce its own hit list.

Thus, taking into account also the definition of precision, a method to compute a *relative precision* (referred to as *RP method*) can be given.

The idea of the RP method is as follows.

If the hits of a meta-search engine are compared to the hits of the search engines used, then a relative precision can be defined for the meta-search engine.

Let

q be a query,

V be the number of hits returned by the meta-search engine under focus, and

T those hits out of these *V* that were ranked by at least one of the search engines used within the first *m* of its hits.

Then, the *relative precision* $RP_{q,m}$ of the meta-search engine is calculated as follows:

$$RP_{q,m} = \frac{T}{V} .$$

$$RP_{q,m} = \frac{T}{V}$$

The value of *m* can be, for example

m = 10 or

m = 5, or

some other value depending on several factors (e.g., the range of the measurement, etc.).

The value of relative precision should be computed for several queries, and an average should be computed.

The *RP* method relies heavily on the hypothesis that the hit lists of search engines contain relevant hits. In other words, the *RP* measure is as good as the hit lists are.

RP METHOD

(Relative Precision of Web meta-search engine)

1. Select meta-search engine to be measured.
2. Define queries q_i, $i = 1,...,n$.
3. Define the value of m; typically $m = 5$ or $m = 10$.
4. Perform searches for every q_i using the meta-search engine

   as well as the search engines used by the meta-search engine, $i = 1,...,n$.

5. Compute relative precision fro q_i as follows:

$$RP_{q_i,m} = \frac{T_i}{V_i}, \quad i = 1,...,n$$

6. Compute average: $\sum_{i=1}^{n} RP_{q_i,m}$

Example

- Let us assume that a meta-search engine uses four search engines.
- Let the query q be „Download ICQ Message Archive”,
- and let us assume further that the meta-search engine returns five hits, i.e., $V = 5$.

Analysing the hit lists of all the search engines:

- the first hit of the meta-search engine is the third on the hit list of the first search engine,
- the second hit was the first in the second search engine,
- the third was the fourth in the third search engine,
- the fourth was the second in the fourth search engine,
- the last one was the third in the second search engine.

⇓

Thus, $T = 5$, and for $m = 10$

the relative precision is $RP_{q,10} = 5 : 5 = 1$.