

## Chapter 8

### Recall, Precision and Accuracy Ratios

Ismaeel Abiodun Sikiru<sup>1</sup>

Mutawakilu Adisa Tihamiyu<sup>2</sup>

<sup>1</sup>*Department of Information and Communication Science,  
University of Ilorin, Nigeria*

<sup>2</sup>*Africa Regional Centre for Information Science, University of  
Ibadan, Nigeria*

#### 8.1 Introduction

The Web has become an ocean of information and resources, which is growing rapidly larger in every microsecond. It has grown from an esoteric system used by a small community of researchers to the current most used system, as an inevitable tool for obtaining information by billions of digital citizens or 'digizens' (Helsper & Eynon, 2009). In a report released by the Internet World Stats (2018) on "World Internet usage and population statistics, June 30, 2018 - update", out of all the seven world regions – Africa, Asia, Europe, Latin America / Caribbean, Middle East, North America, and Oceanic / Australia, Africa has the least Internet penetration rate of 36.1% of the total population, but with a growth of 10,199% in the period 2000 to 2018, against North America with the highest penetration rate of 95%.

The Web is both a huge database of web pages, as well as a gateway through which the search of information systems or databases of various organizations is conducted. Such information systems or databases include those of online stores like Amazon and Jumia, app stores like Google Play or Samsung Galaxy, as well as the online public access catalogues (OPACs) of libraries. Traditionally, people often directly search the information systems or databases of organizations wherein, they are staff, students, customers or permitted visitors. On the web however, search engines such as Google or Yahoo often do the searching of these other systems and databases for people, thereby saving them from having to search these other systems themselves.

Due to the dividends of Web in this 21st century, many people have never encountered, and thus even have no interest in the issues and challenges of retrieving information from such databases, despite its importance (Oppenheim, Morris, McKnight, & Lowley, 2000). However, all information found on the Web through search engines or directly from other information systems or databases usually needs to be evaluated and filtered, as it may include plenty of irrelevant information to the subject at hand.

Regrettably, a Web surfer may not be aware of many available search engines that can be used to get information on a topic, which by the effect, drives him/her to use different search strategies, some of which might not be effective or efficient (Kumar and Prakash, 2009). While in the course of searching the Web, other information systems or databases, there are some notable challenges that a searcher faces which need to be answered at the initial searching process. The challenges are listed here in an interrogative approach:

- (a) Which search engine or database would get me quickly the best search results that include only or mostly relevant information and also exclude all or most of the irrelevant information?
- (b) What search expression (or query) comprising important words, terms, names, URLs, etc.) best describes my information need that I should input to the search engine or database?
- (c) How can I determine quickly which items in the initial search results provided by the search engine are most relevant to my needs?
- (d) How many items or pages of the search results should I look at before determining if the results are excellent, good, fair, poor or adequate for my needs?
- (e) If the initial search results are not adequate, how do I revise or refine my initial query to get better subsequent search results?
- (f) When should I end the search, satisfied or frustrated?

From the above questions, searching is obviously a multistep and iterative process. As an iterative process, an initial query may not be properly worded, and therefore needs to be improved using new words and terms identified from the

previous search result(s). In an attempt to obtain improvement in the relevance of items in the final result, search process expects from the Web searcher the following nine steps (Figure 8.1):

1. Define your information need (also known as request). Describe the information need as precisely as possible with concepts (i.e. Keywords or sentences)
2. Choose an appropriate information resource (search engine database, full text or bibliographic database, library catalogue, document repository, the Web)
3. Choose an appropriate search tool (e.g. Google, Yahoo, Ask, or as provided by a databases or information system).
4. Identify and list relevant search terms derived from your search request that you would include in your query for use in searching.
5. Modify the search terms to suit the chosen information resource (i.e. use the vocabulary dictionary of each information resource to get equivalent or other terms used by the resource)
6. Combine the original, modified or augmented search terms to create a search query you intend to begin your search with.
7. Run an initial search on the information resource, by entering your search query in the search box of the search engine or tool.
8. Evaluate your initial search to determine how good the search result is, by examining and comparing the retrieved items in the search result with your information need.
9. Modify your search query based on the previous results and run new searches as necessary until you are satisfied with the results you have obtained.
10. If your answers to the evaluation questions continue to be 'No' after repeatedly modifying your search query, then you may have to loop back to any of the earlier steps before the 'Enter Search Query' step.

You should bear in mind the following points. Firstly, step 6 above requires that you evaluate each search result, usually while still working with the search engine, in order to determine how good each search result is overall, and to also determine

which items in each search result you need to copy, paste and save in Step 8. Secondly, the quality of the search results provides by a search engine or information system depends critically on (a) what you yourself do in steps 1 to 6, as well as how good the search engine is in matching terms in your search queries with words, terms, and phrases in its database. Often, because search engines and their databases have been researched and built to be effective and efficient for searchers as much as possible, the quality of what you get usually depends on you!

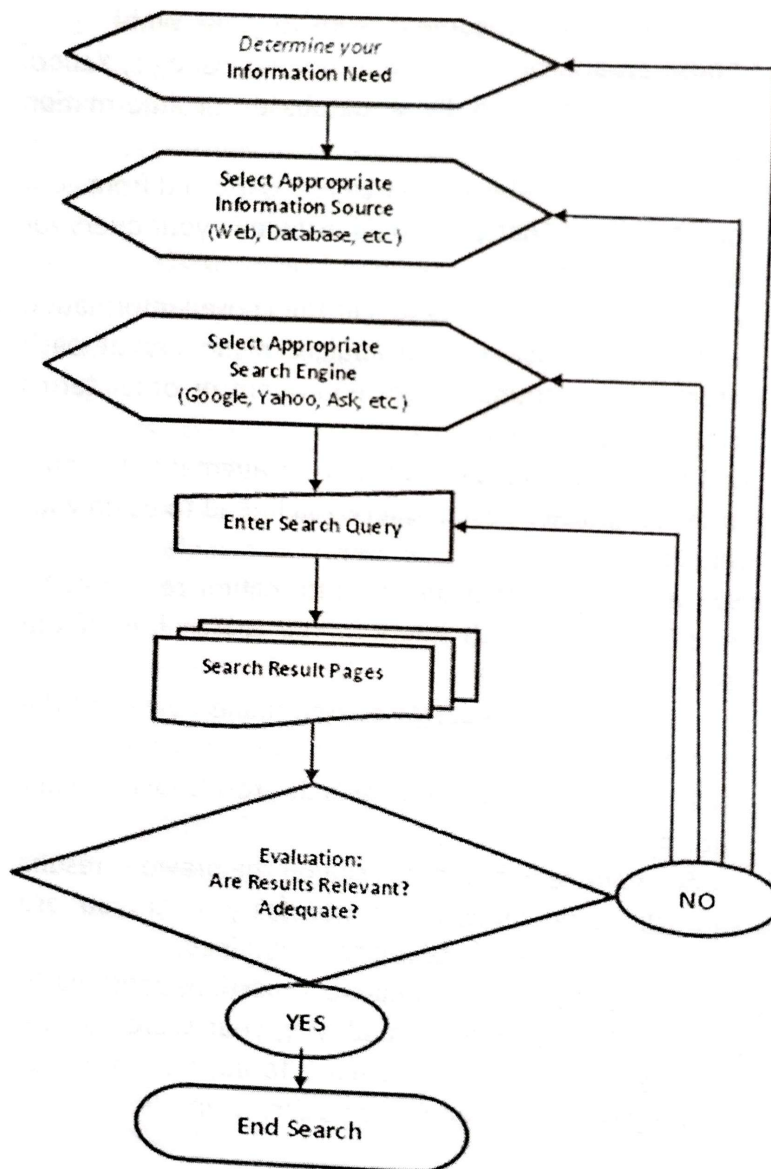


Figure 8.1: Iterative Search Process.

Let us start from the beginning of the search process by considering step 1 above, which requires that you define or describe your information need with appropriate words simply, precisely and adequately. For example, consider an information need implied by this question: *What are the effects of e-books on tertiary education students?* Five different key words or concepts can be drawn out from the question which are: *effects, e-books, tertiary, education, students*. Then, in step 2 you need to identify other terms that are synonyms of the initial concepts, as shown in Table 1. In step 3, you need to find out the terms actually used by the chosen information system, which might or might not be the same as the initial or synonym concepts. Usually, the best search results are obtained when a searcher uses the same terms to search as the terms that were used by a search engine or information system when it was indexing its resources. Finally, steps 4 to 7 also depend on you – how you combine the concepts in the initial and subsequent queries, and how you evaluate the initial and subsequent search results. In a nutshell, in order to obtain the best search results, all the steps in the search process must be well conducted.

Table 8.1: Keywords (concepts) in query and their approximate synonyms.

|                                                       | <b>Concept 1</b> | <b>Concept 2</b> | <b>Concept 3</b>   | <b>Concept 4</b> | <b>Concept 5</b> |
|-------------------------------------------------------|------------------|------------------|--------------------|------------------|------------------|
| <i>Query Concepts</i>                                 | effects          | e-books          | tertiary           | education        | students         |
| <i>Near synonyms of each query concept</i>            | influence        | electronic books | collegiate         | training         |                  |
|                                                       | impact           | digital books    | higher education   |                  |                  |
|                                                       | consequence      | online books     | Post-secondary     |                  |                  |
| <i>Concept used in database, document or web page</i> | impact           | online books     | tertiary education |                  | students         |

Although what you get usually depends on you, various yardsticks or metrics have been researched and recommended for use to evaluate the search results performance of search engines and other information systems. The rest of this chapter examines and explain the most common of these metrics.

## 8.2 Recall and Precision ratios

The earliest suggested and most commonly mentioned yardsticks are known as recall and precision ratios.

- **Recall** is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage.
- **Precision** is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage.

### 8.2.1 Real Life Illustration of Recall and Precision

A simple yet good illustration of the ideas of recall and precision is the following possible real life usages of recall and precision:

Imagine that, your friend gave you a birthday surprise every year in the last 10 years. *However, one day, your friend asks you "Buddy, do you remember all the birthday surprises from me?"* This simple question is likely to be tough to answer because you need to recall all 10 surprising events from your memory.

Now, with the aid of recall and precision metrics approach, the solution is as follows:

Let us suppose your girlfriend has a particular set of 10 surprises in her mind which is what she expected to be told as her information need. Recall that, Recall ratio is the number of events (surprises) you can *correctly* recall divided by the number of all the correct events (that she expects you to recall). It measures effectiveness of recall! So, (1) if you can recall all 10 events correctly, then, your recall ratio is  $10 / 10 = 1.0$  (or 100%), while (2) if you can recall only 7 events correctly, your recall ratio is  $7 / 10 = 0.7$  (70%).

Precision ratio is the number of events you recalled *correctly* divided by the number of all events you are able to recall, which usually include *both correct and incorrect events*. In other words, the precision ratio measures how

precise and efficient your recall efforts are. Suppose in example (1) above you made exactly 10 attempts in getting the 10 correct events. Then your precision ratio is 10 correct recalls divided by 10 recall attempts, which is also 1.0 (100%). However, in example (2), you also made 10 recall attempts, but got only 7 correct recalls. So, the preciseness or efficiency of recalling the events is the 7 correct recalls divided by 10 recall attempts, which is 0.7 (70%).

Next, suppose you can actually recall many events (some of which are) correctly in the last ten years, while the others were not. Suppose you eventually told her 16 events in 16 recall attempts, out of which only 8 events are among the particular 10 events she has in mind. In that case, your recall ratio is the 8 correct events out of the 10 she has in mind, which is  $8 / 10 = 0.8$  (80%). Your recall ratio improved by 10%, but only after six more attempts beyond scenario (2) above. You improved your recall ratio by 10%, which means that your effectiveness in recalling correct events improved by 10%, but at the cost of 6 (60%) more attempts.

Would you say you are becoming more precise or efficient in example (3)? Actually your precision ratio in example (3) is only 8 correct events out of 16 recalled events, which is only  $8 / 16 = 0.5$  (50%). Your recall ratio improved by 10%, but your precision ratio decreased by 20%. So you have become more effective at recalling correctly, but less efficient in doing so!

### 8.2.2 Recall ratio vs. Precision ratio

The above examples illustrated a natural inverse relationship between the concepts and measurement of recall and precision ratios of the search results provided by search engines in response to user queries. This can be shown both mathematically and graphically. Mathematically,

$$Recall = \alpha \frac{1}{Precision}$$

Which you can and should confirm this using the recall and precision ratios calculated in examples (1) to (3) above. Recall ratios range from 0 to 1, likewise precision ratios. The inverse or trade off relationship that exists naturally between them for

search results provided by an information system in response to different queries that users provide is illustrated in Figure 2.

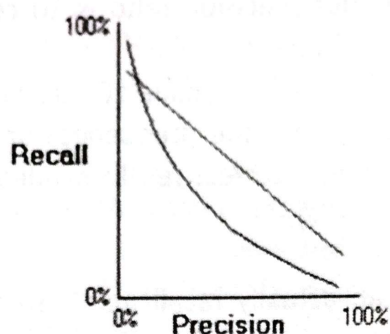


Figure 8.2: Recall – Precision relationship

In the figure, the two distinct lines may represent the recall-precision graphs of two search engines or information systems. While the exact slope of the curve may vary between systems, the general inverse relationship between recall and precision remains for every information system.

Consequently, a system can increase its ability to recall by returning more documents; because the recall ratio is a non-decreasing function of the number of documents retrieved. A system that returns all documents in its database for a query will surely have 100% recall of all the relevant items in its database, but the precision ratio of such response to the query will be very low, due to the likely higher number of non-relevant items returned along with the relevant items. The converse is also true, as it is possible for a system to aim for high precision; but at the cost of very low recall of relevant items from its database. This naturally occurring inverse relationship between precision and recall ratios forces information systems designed for general use to go for compromise between them. But, in real life, some information search tasks particularly need good precision, whereas others need good recall.

### 8.3 F score - A combined measure

A combined measure that measures simultaneously the recall (R) and Precision (P) performance of the search results from an information system is the *F score* (weighted harmonic mean). The F-score is a measure which combines both recall and

precision measures using a weighting factor  $\alpha$ , where high  $\alpha$  means that precision is more important.

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad \dots (1)$$

The harmonic mean is a very conservative average. People usually use the balanced F1 measure, i.e.

$$\text{with } \beta = 1 \text{ (thereby, } \alpha = \frac{1}{2}) \quad \dots (2)$$

Applying (2) in (1), we have:

$$F = \frac{2PR}{(P+R)} \quad \dots (3)$$

The notions of recall and precision were first used by Kent (1971), although the term precision did not appear until later. The F measure (or, rather its complement  $E = 1 - F$ ) was provided by van Rijsbergen (1979). He provided an extensive theoretical discussion to show that the principle of decreasing marginal relevance would at some point make a searcher for information to be unwilling to sacrifice a unit of precision for an added unit of recall and vice versa. He then proposed the F measure (which is the harmonic mean of the recall and precision ratios) as the appropriate method for combining precision and recall.

#### 8.4 Relevance

The concept of relevance refers to the extent of usefulness of the documents retrieved during a search to meeting the information need that necessitated the search. Relevance is assessed relative to an information need, not a query was used for a search. But an information need might differ very much from a query used in a search. For example, an information need of a secondary school physics teacher might be getting accurate information on the "strengths and weaknesses of using different teaching methods or aids in the teaching of physics at the secondary school level". This information need might be translated into a query such as:

*teaching AND physics AND "secondary school" AND (method OR aid)*

A document is said to be relevant if it helps in meeting the stated information need, not because it just happens to contain some of or all the words in the query. This distinction is often misunderstood in practice, because the words and sentences that describe an information need are usually in a searcher's mind and often not written down. Nevertheless, an information need is present in a searcher's mind which drives him to do a search.

Searching involves selecting and combining words which may be or may not be those in the searcher's mind to form a search query. Moreover, words can have different meanings. For instance, if a user types the word *python* into a web search engine, he might intend to get information on where to purchase a pet python, or information on how to learn the Python programming language, or what and how pythons eat. So just typing "python" into search box of search engine would often retrieve so many documents or web pages most of which would not be relevant to the information need. It is by this effect, usually very difficult for a system to know what the precise information need is, when such a one word query is used. Nevertheless, the searcher can judge the quality of the returned results on the basis of their relevance to his information need expressed with just that one word. Therefore, to evaluate a system, an overt expression of an information need in the form of one or more words is required, which can be used for judging returned documents as *relevant* or *non-relevant*.

Relevance of returned documents can reasonably be thought of as a scale such as ranging from zero, through very low, low, moderate, high, very high, etc. However, in assessing relevance for the purposes of calculating recall and precision ratios or F score, a simplification is often made that, each document in the database being searched is totally relevant to the information need or totally irrelevant to the need. Thus, each document in the search results is also totally relevant or totally irrelevant to the need. This assertion is illustrated with the diagram below, where:

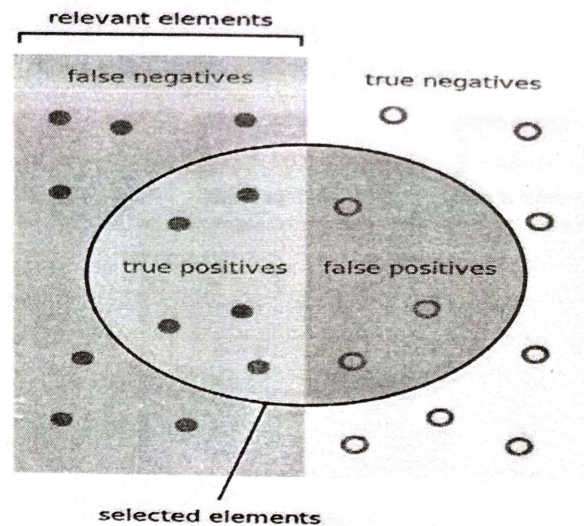


Figure 8.3: States of Information search query

- i. The database of documents being searched is represented with the largest rectangle, which is divided into two smaller rectangles representing documents that are *relevant* and documents that are *not relevant*.
- ii. The search results is represented with the oval, which is also divided into two parts representing documents that are relevant and retrieved (*true positives*) or not relevant but retrieved (*false positives*).
- iii. The documents not retrieved from the database also divides into two: those relevant but not retrieved (*false negatives*) and those not relevant and not retrieved (*true negatives*)

Using the above explanation of the meaning of the relevance of documents to a particular information need, and the above four sets of documents, let's analyse relevance based on binary decision. Mathematically,

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

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

N.B:

- All Retrieved documents are classified by the retrieval systems Positives (*vertically pink shaded*)
- The Relevant documents are the really **Positives** (*horizontally yellow shaded*)
- Relevant and Retrieved are the **True Positives** (*intersection of the pink and yellow shaded*)

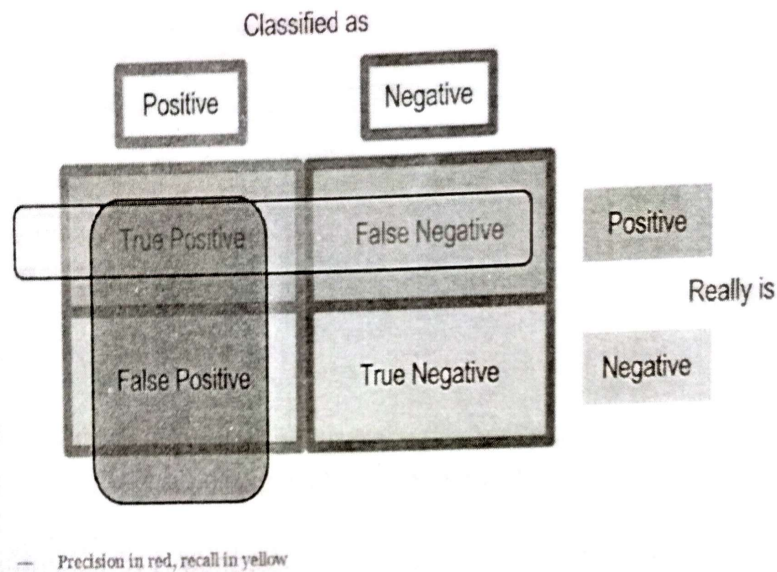


Figure 8.4: Classification of documents during retrieval.

## 8.5 Accuracy

Accuracy is another concept that has been mentioned as possibly useful yardstick for assessing the ability of an information retrieval system (such as search engine or database) to simultaneously correctly classify documents relevant and not relevant types. The number of such correct classifications is then expressed as a ratio of the total number of documents in the database. Using the TP, TN, FP, and FN classification of documents as illustrated in the diagram above, the **accuracy index** can be stated as:

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

However, accuracy has been seen in literature as not a useful measure for assessing the effectiveness and efficiency of retrieval in large databases, such as in web information retrieval. This is because the focus in searching is usually on relevant documents, which is represented by TP, but in a large database TP is almost always a tiny fraction of all the documents in a database, and so is dwarfed by TN. Thus, even if good information retrieval system (IRS) retrieves only relevant documents in response to a query, the accuracy index of the system measured by the above formula will not be much

different from that of a bad system that does not do well at all. Hence, the accuracy index cannot help us evaluate or compare IR systems (Schellekens, 2012).

## 8.6 Ranking of retrieved documents

When a search engine retrieves so many supposedly relevant documents in response to a query, some ranking of the documents is critical so that documents can be presented to the searcher in the rank order of their estimated relevance to the information need. Relevance ranking of the retrieved documents is usually done automatically by most search engines based on such indices as:

- **Term Frequency (TF):** This is the frequency of occurrence of each query keyword in a document. Higher frequency for a query keyword gives the keyword more importance in the document.
- **Inverse Document Frequency (IDF):** Number of documents in the database in which each query keyword occurs. The lower the number of documents the more important they keyword is considered to be.
- **Hyperlinks to documents:** The number of hyperlinks that point to a document from other documents in the database. The more links pointing to a document, the more important the document is considered to be.

### (a) Relevance ranking using Term Frequency (TF) and Inverse Document Frequency (IDF):

The number of documents relevant to a query can be enormous if only term frequencies (TF) is used as basis for retrieving documents. Thus, in finally deciding which documents to retrieve,

- Words that occur in title, author list, section headings, etc. are given greater importance, while words whose first occurrence is late in the document are given lower importance.
- Stop words such as "a", "an", "the", "it" are not considered.

- If the query keywords occur close together in a document, the document is given more importance than if the keywords occur far apart.
- Documents are returned in decreasing order of relevance score, with usually only top few documents returned.

Therefore, the ranking of documents using TF and IDF is actually performed as follows:

- (1) Let  $t$  be a term in query  $Q$
- (2) Let  $n(d)$  be the number of terms in document  $d$
- (3) Let  $n(d, t)$  be the number of occurrences of each term  $t$  in each document  $d$
- (4) The relevance of a document  $d$  to a term  $t$  is:

$$TF(d, t) = \log \left( 1 + \frac{n(d, t)}{n(d)} \right)$$

- (5) The log factor is then used to avoid excessive weight to the most frequent terms, such that the relevance of document  $d$  to term  $t$  is calculated as:

$$r(d, Q) = \sum_{t \in Q} \frac{TF(d, t)}{n(t)}$$

#### **(b) Relevance ranking based on hyperlinks**

This method is connected to social networking theories that explain and rank the prestige of persons in terms of the number and status of the connections they have with others. For instance a personality like the President of Nigeria is expected to have very high prestige because he/she is known by (a) many (b) prestigious other people. An example of ranking methods based on hyperlinks is Hub and Authority Based Ranking which can be used by a search engine to rank retrieved web pages. In this method:

- A hub is a page that stores links to many pages (on a topic)
- An authority is a page that contains actual information on a topic
- Each retrieved web page gets a hub prestige score based on prestige of authorities that it points to.
- Each page gets an authority prestige based on prestige of hubs that point to it.

## 8.7 Calculating of recall, precision and other scores

Two Information retrieval systems, A and B, are being compared. Both are given the same query, applied to a collection of 1000 documents. System A returns 420 documents, of which 50 are relevant to the query. System B returns 90 documents, of which 25 are relevant to the query. Within the whole collection there are in fact 80 documents relevant to the query.

Create a contingency table of the results for each system, and compute the following:

- Recall
- Precision
- Accuracy
- F score

### Solution

| System A     | Relevant | Non-relevant | Total |
|--------------|----------|--------------|-------|
| Returned     | 50       | 370          | 420   |
| Not returned | 30       | 550          | 583   |
| Total        | 80       | 920          | 1000  |

| System B     | Relevant | Non-relevant | Total |
|--------------|----------|--------------|-------|
| Returned     | 25       | 65           | 90    |
| Not returned | 55       | 855          | 910   |
| Total        | 80       | 920          | 1000  |

### System A:

$$\text{System A's Recall} = \frac{TP}{TP+FN} = \frac{50}{80} = 0.625 = 62.5\%$$

$$\text{System A's Precision} = \frac{TP}{TP+FP} = \frac{50}{420} = 0.119 = 11.9\%$$

$$\begin{aligned}\text{System A's Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN} = \frac{50+550}{50+550+370+30} \\ &= \frac{600}{1000} = 0.6 = 60\%\end{aligned}$$

$$\text{System A's F Score} = \frac{2PR}{(P+R)} = \frac{2 \times 0.119 \times 0.625}{(0.119+0.625)} = \frac{0.149}{0.744} = 0.2$$

### System B:

$$\text{System B's Recall} = \frac{TP}{TP+FN} = \frac{25}{80} = 0.313 = 31.3\%$$

$$\text{System B's Precision} = \frac{TP}{TP+FP} = \frac{25}{90} = 0.278 = 27.8\%$$

$$\begin{aligned}\text{System B's Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN} = \frac{25+855}{25+855+65+55} \\ &= \frac{880}{1000} = 0.88 = 88\%\end{aligned}$$

$$\text{System B's F Score} = \frac{2PR}{(P+R)} = \frac{2 \times 0.278 \times 0.313}{(0.278+0.313)} = \frac{0.174}{0.591} = 0.294$$

## 8.8 Conclusion

Recall and precision are the two fundamental IR evaluation metrics. Both are the cornerstones for many other developed measures and are easy to understand by all users. To the practitioner's view, these two evaluation measures are interesting because they lead to more intuitive interpretations like, how much time people are reading useless documents (low precision), or how many relevant documents they are missing (low recall). Both precision and recall need to be taken into account when evaluating retrieval systems. It is not sufficient to pick one and use only that, because dependence on just one out of precision and recall can lead to extreme and unhelpful solutions. For example, a system that returns every document indiscriminately will have 100% recall; while another system that returns only a single correct document is 100% precise. The former system is of little or no help because the searcher needs to sift through the massive documents returned to identify those documents that are relevant to needs. The latter system is not much better, as the only one document it returns is unlikely to satisfy the need.

## References

- Helsper, E. & Eynon, R. (2009) Digital natives: where is the evidence? *British Educational Research Journal*. pp. 1-18. DOI: 10.1080/01411920902989227
- Jurafsky, D. & Manning, C. (Producers), (2012). Evaluation of text classification - Precision, Recall, and the F measure

[MP4]. Available from  
<https://www.youtube.com/watch?v=2akd6uwtowc>.  
Stanford NLP.

- Kent, A. (1971). *Information Analysis and Retrieval*, 3rd ed. New York, Becker and Heys.
- Lavrenko, V. (Producer), (2014). *Evaluation: building blocks – Precision and Recall* [MP4]. Available from  
<https://www.youtube.com/watch?v=mctizdBujk4>
- Lipani, A. (2016). *Fairness in Information Retrieval*. In: *Proceedings of SIGIR'16 Annual Conference*, July 17 – 21, 2016.
- Oppenheim, C., Moris, A, Mcknight, C., & Lowley, S. (2000). *The evaluation of WWW search engines. Journal of Documentation*, 56 (2), 190-211.
- Sampath Kumar B. T. & Prakash, J. N. (2009). *Precision and Relative Recall of Search Engines: A Comparative Study of Google and Yahoo. Singapore Journal of Library & Information Management*, Volume 38.
- Schellekens, M. (2012). *Information Retrieval Evaluation*. TA: Ang Gao, University College Cork.
- Silberschatz, K. and Sudarshan, I. (2005). *Database System Concepts – 5th edition*.
- Van Rijsbergen, C.J. (1979). *Information Retrieval*, Butterworths, London, 2nd edition.  
[https://www.creighton.edu/fileadmin/user/HSL/docs/ref/Searching\\_-\\_Recall\\_Precision.pdf](https://www.creighton.edu/fileadmin/user/HSL/docs/ref/Searching_-_Recall_Precision.pdf)  
<https://www.quora.com/What-is-the-best-way-to-understand-the-terms-precision-and-recall>