Assignment No B-1

TITLE: Precision and Recall

AIM: Implement a program to calculate precision and recall for sample input. (Answer set A, Query q1, Relevant documents to query q1- Rq1)

OBJECTIVE: To study

- Performance Evaluation,
- Retrieval Performance Evaluation
- Precision and Recall.

PROBLEM STATEMENT:-

Write a program to implement a program to calculate precision and recall for sample input. (Answer set A, Query q1, Relevant documents to query q1- Rq1)

TOOLS / ENVIRONMENT:

- **S/W:**
  - Can be developed on any platform Windows /Linux.
  - C /C++/Java Language
- **H/W:**
  - Any basic configuration loaded machine (e.g. P IV)

THEORY:

Introduction

Before the final implementation of an information retrieval system, an evaluation of the system is usually carried out. The type of evaluation to be considered depends on the objectives of the retrieval system. Clearly, any software system has to provide the functionality it was conceived for. Thus, the first type of evaluation which should be considered is a functional analysis in which the specified system functionalities are tested one by one. Such an analysis should also include an error analysis phase in which, instead of looking for functionalities, one behaves erratically trying to make the system fail. It is a simple procedure which can be quite useful for catching programming errors. Given that the system has passed the functional analysis phase, one should proceed to evaluate the performance of the system.

The most common measures of system performance are time and space. The shorter the response time, the smaller the space used, the better the system is considered to be. There is an inherent tradeoff between space complexity and time complexity which frequently allows trading one for the other.

In a system designed for providing data retrieval, the response time and the space required are usually the metrics of most interest and the ones normally adopted for evaluating the system. In this case, we look for the performance of the indexing structures (which are in place to accelerate the search), the interaction with the operating system, the delays in communication channels, and the overheads introduced by the many software layers which are usually present. We refer to such a form of evaluation simply as **performance evaluation**.

In a system designed for providing information retrieval, other metrics, besides time and space, are also of interest. In fact, since the user query request is inherently vague, the retrieved documents are not exact answers and have to be ranked according to their relevance to the query. Such relevance ranking introduces a component which is not present in data retrieval systems and which plays a central role in information retrieval. Thus, information retrieval systems require the evaluation of how precise is the answer set. This type of evaluation is referred to as **retrieval performance evaluation**.

Retrieval Performance Evaluation

When considering retrieval performance evaluation, we should first consider the retrieval task that is to be evaluated. For instance, the retrieval task could consist simply of a query processed in batch mode (i.e., the user submits a query and receives an answer back) or of a whole interactive session (i.e., the user specifies his information need through a series of interactive steps with the system). Further, the retrieval task could also comprise a combination of these two strategies. Batch and interactive query tasks are quite distinct processes and thus their evaluations are also distinct. In fact, in an interactive session, user effort, characteristics of the interface design, guidance provided by the system, and duration of the session are critical aspects which should be observed and measured. In a batch session, none of these aspects is nearly as important as the quality of the answer set generated.

Besides the nature of the query request, one has also to consider the setting where the evaluation will take place and the type of interface used. Regarding the setting, evaluation of experiments performed in a laboratory might be quite distinct from evaluation of experiments carried out in a real life situation. Regarding the type of interface, while early bibliographic systems present the user with interfaces which normally operate in batch mode, newer systems (which are been popularized by the high quality graphic displays available nowadays) usually present the user with complex interfaces which often operate interactively.

Retrieval performance evaluation in the early days of computer-based information retrieval systems focused primarily on laboratory experiments designed for batch interfaces. In the 1990s, a lot more attention has been paid to the evaluation of real life experiments. Despite this tendency, laboratory experimentation is still dominant. Two main reasons are the repeatability and the scalability provided by the closed setting of a laboratory.

Recall and Precision

Consider an example information request I (of a test reference collection) and its set R of relevant documents. Let IRI be the number of documents in this set. Assume that a given retrieval strategy (which is being evaluated) processes the information request I and generates a document answer set A. Let IAI be the number of documents in this set. Further, let IRaI be the number of documents in the intersection of the sets R and A. Following figure illustrates these sets.

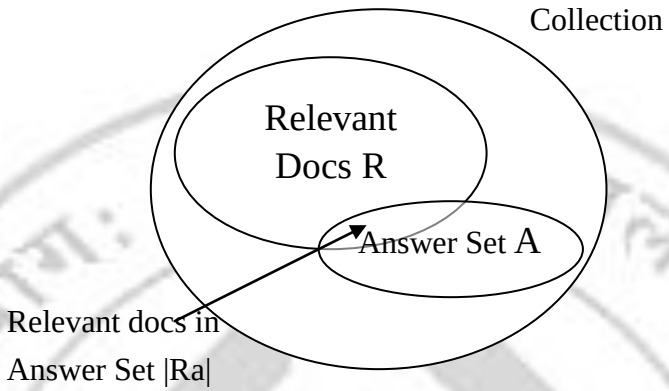
The recall and precision measures are defined as follows.

Recall is the fraction of the relevant documents (the set R) which has been retrieved i.e.,

$$\text{Recall} = \frac{|\,Ra\,|}{|\,R\,|}$$

Precision is the fraction of the retrieved documents (the set A) which is relevant i.e.,

$$\text{Pr}\ ecision = \frac{|\,Ra\,|}{|\,A\,|}$$



Recall and precision, as defined above, assume that all the documents in the answer set A have been examined (or seen). However, the user is not usually presented with all the documents in the answer set A at once. Instead, the documents in A are first sorted according to a degree of relevance (i.e., a ranking is generated). The user then examines this ranked list starting from the top document. In this situation, the recall and precision measures vary as the user proceeds with his examination of the answer set A. Thus, proper evaluation requires plotting a precision versus recall curve as follows.

As before, consider a reference collection and its set of example information requests. Let us focus on a given example information request for which a query q is formulated. Assume that a set Rq containing the relevant documents for q has been defined. Without loss of generality, assume further that the set Rq is composed of the following documents

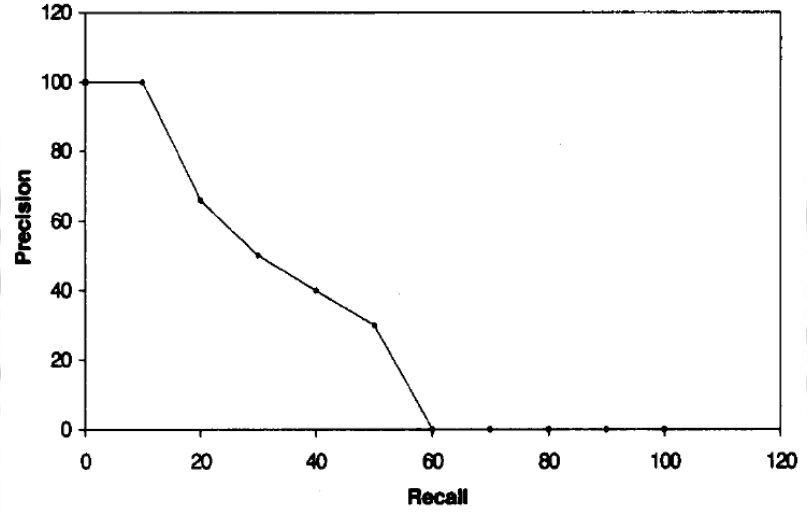
- $R_{q1} = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$

Thus, according to a group of specialists, there are ten documents which are relevant to the query q. Consider now a new retrieval algorithm which has just been designed. Assume that this algorithm returns, for the query q, a ranking of the documents in the answer set as follows.

Ranking for query q:

01. d_{123} •	06. d_9 •	11. d_{38}
02. d_{84}	07. d_{511}	12. d_{48}
03. d_{56} •	08. d_{129}	13. d_{250}
04. d_6	09. d_{187}	14. d_{113}
05. d_8	10. d_{25} •	15. d_3 •

The documents that are relevant to the query q are marked with a bullet after the document number. If we examine this ranking, starting from the top document, we observe the following points. First, the document d123 which is ranked as number 1 is relevant. Further, this document corresponds to 10% of all the relevant documents in the set Rq • Thus, we say that we have a precision of 100% at 10% recall. Second, the document d56 which is ranked as number 3 is the next relevant document. At this point, we say that we have a precision of roughly 66% (two documents out of three are relevant) at 20% recall (two of the ten relevant documents have been seen). Third, if we proceed with our examination of the ranking generated we can plot a curve of precision versus recall as illustrated in figure below.



[Figure: Precision at 11 standard recall levels.]

The precision at levels of recall higher than 50% drops to 0 because not all relevant documents have been retrieved. This precision versus recall curve is usually based on 11 (instead of ten) standard recall levels which are 0%, 10%, 20%, ... , 100%. For the recall level 0%, the precision is obtained through an interpolation procedure as detailed below.

In the above example, the precision and recall figures are for a single query. Usually, however, retrieval algorithms are evaluated by running them for several distinct queries. In this case, for each query a distinct precision versus recall curve is generated. To evaluate the retrieval performance of an algorithm over all test queries, we average the precision figures at each recall level as follows.

- $\overline{P}(r_j) = \sum_{i=1}^{N_q} \frac{P_i(r_j)}{N_q}$
- where
  - $\overline{P}(r_j)$ is the average precision at the recall level r_j
  - $P_i(r_j)$ is the precision at recall level r_j for the *i*-th query

Since the recall levels for each query might be distinct from the 11 standard recall levels, utilization of an interpolation procedure is often necessary. For instance, consider again the set of 15 ranked documents presented above. Assume that the set of relevant documents for the query q has changed and is now given by

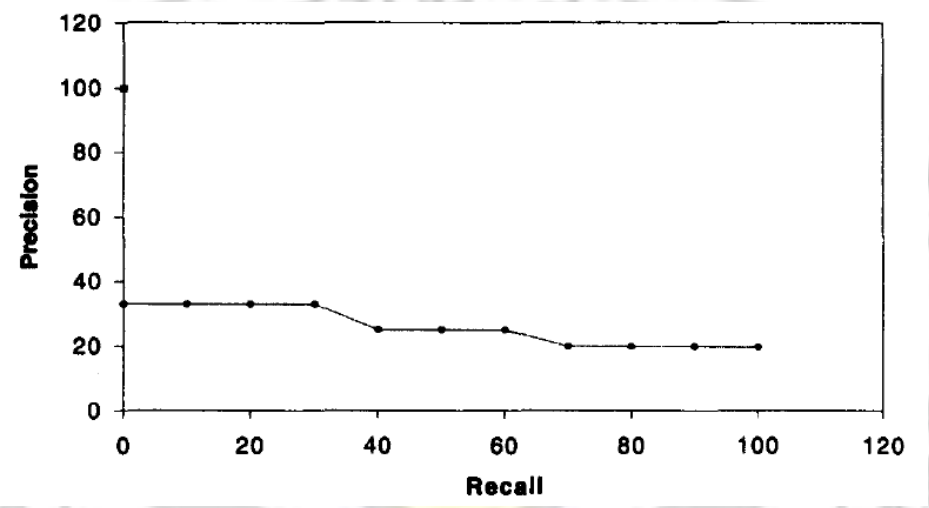
$R_q = \{d_3, d_{56}, d_{129}\}$

In this case, the first relevant document in the ranking for query q is dS6 which provides a recall level of 33.3% (with precision also equal to 33.3%) because, at this point, one-third of all relevant documents have already been seen. The second relevant document is d129 which provides a recall level of 66.6% (with precision equal to 25%). The third relevant document is d3 which provides a recall level of 100% (with precision equal to 20%). The precision figures at the 11 standard recall levels are interpolated as follows.

Let r_j , $j \in \{1, 2, 3, \ldots, 10\}$, be a reference to the *j*-th (i.e., rs is a reference to the recall level 50%). Then, $P(rj) = max_{\forall r | r_j \leq r} P(r)$

which states that the interpolated precision at the j-th standard recall level is the maximum known precision at any recall level between the j-th recall level and the (j +1)-th recall level.

In our last example, this interpolation rule yields the precision and recall figures illustrated in figure below.

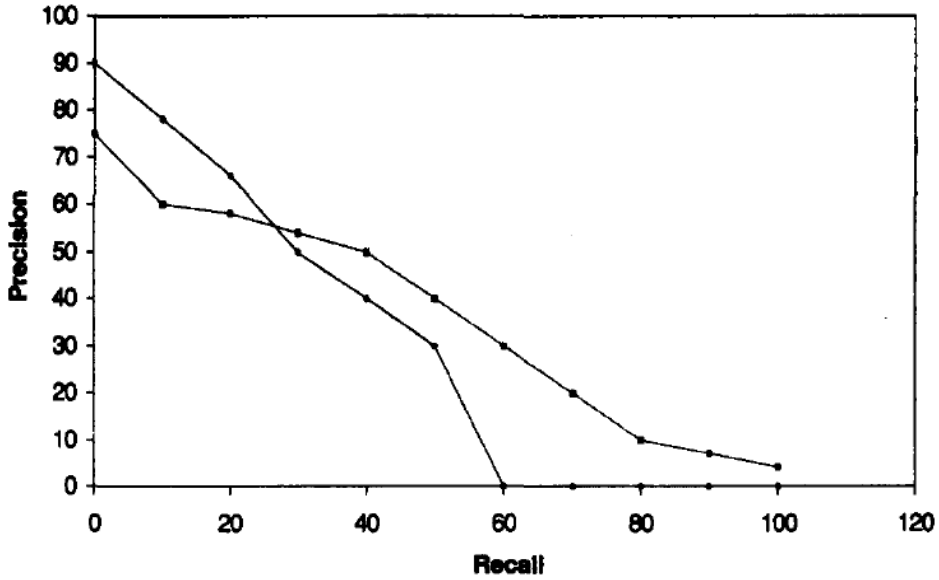


[Figure: Interpolated precision at 11 standard recall levels relative to $R_q = \{d_3 , d_{56} , d_{129}\}$]

At recall levels 0%, 10%, 20%, and 30%, the interpolated precision is equal to 33.3% (which is the known precision at the recall level 33.3%). At recall levels 40%, 50%, and 60%, the interpolated precision is 25% (which is the precision at the recall level 66.6%). At recall levels 70%, 80%, 90%, and 100%, the interpolated precision is 20% (which is the precision at recall level 100%).

The curve of precision versus recall which results from averaging the results for various queries is usually referred to as precision versus recall figures. Such average figures are normally used to compare the retrieval performance of distinct retrieval algorithms. For instance, one could compare the retrieval performance of a newly proposed retrieval algorithm with the retrieval performance of the classic vector space model. Figure 3.4 illustrates average precision versus recall figures for two distinct retrieval algorithms. In this case, one algorithm has higher precision at lower recall levels while the second algorithm is superior at higher recall levels.

One additional approach is to compute average precision at given document cutoff values. For instance, we can compute the average precision when 5, 10, 15, 20, 30, 50, or 100 relevant documents have been seen. The procedure is analogous to the computation of average precision at 11 standard recall levels but provides additional information on the retrieval performance of the ranking algorithm.



[Figure Average recall versus precision figures for two distinct retrieval algorithms.]

Average precision versus recall figures are now a standard evaluation strategy for information retrieval systems and are used extensively in the information retrieval literature. They are useful because they allow us to evaluate quantitatively both the quality of the overall answer set and the breadth of the retrieval algorithm. Further, they are simple, intuitive, and can be combined in a single curve. However, precision versus recall figures also have their disadvantages and their widespread usage has been criticized in the literature. We return to this point later on. Before that, let us discuss techniques for summarizing precision versus recall figures by a single numerical value.

IMPLEMENTING THE SOLUTION:

$R_{q1} = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$

|Ra|=10

string Rq[10] = {"d3", "d5", "d9", "d25", "d39", "d44", "d56", "d71", "d89", "d123"};

Ranking for query q:

01. d_{123} •	06. d_9 •	11. d_{38}
02. d_{84}	07. d_{511}	12. d_{48}
03. d_{56} •	08. d_{129}	13. d_{250}
04. d_6	09. d_{187}	14. d_{113}
05. d_8	10. d_{25} •	15. d_3 •

string A[15] = {"d123", "d84", "d56", "d6", "d8", "d9", "d511", "d129", "d187", "d25", "d38", "d48", "d250", "d113", "d3"};

OUTCOME:

Documents	**|Ra|**	**|A|**	**Precision=Ra /A**	**Recall= Ra /R**
d123	1	1	100%	10%
D123,d84	1	2	50	10
D123,d84,**d56**	2	3	66	20
D123,d84,**d56**,d6	2	4	50	20
D123,d84,**d56**,d6,d8	2	5	40	20
D123,d84,**d56**,d6,d8,**d9**	3	6	50	30
D123,d84,**d56**,d6,d8,**d9**,d511	3	7	42.85	30
D123,d84,**d56**,d6,d8,**d9**,d511,d129	3	8	37.5	30
D123,d84,**d56**,d6,d8,**d9**,d511,d129,d187	3	9	33.33	30
D123,d84,**d56**,d6,d8,**d9**,d511,d129,d187,**d25**	4	10	40	40
D123,d84,**d56**,d6,d8,**d9**,d511,d129,d187,**d25,**d38	4	11	36.36	40
D123,d84,**d56**,d6,d8,**d9**,d511,d129,d187,**d25,**d38,d48	4	12	33	40
D123,d84,**d56**,d6,d8,**d9**,d511,d129,d187,**d25,**d38,d48,d250	4	13	30.76	40
D123,d84,**d56**,d6,d8,**d9**,d511,d129,d187,**d25,**d38,d48,d250, d113	4	14	28.57	40

(Students should write here the implementation for their program. Students should attach printout of their programs with commands used to run the programs. Also attach the proper outputs of programs.)

CONCLUSION: Implementation is concluded by stating precision and recall for sample input.

REFERENCES:

- T1: Ricardo Baeza-Yates , Berthier Riberio Neto, Modern Information Retrieval, Pearson Education, ISBN:81-297-0274-6.
- T2. C.J. Rijsbergen, Information Retrieval, (www.dcs.gla.ac.uk), Second Edition ISBN:978-408709293

FAQ:

1. What do you meant by performance evaluation?
2. What are different measurable quantities consider for performance evaluation?
3. On which parameters the performance of system is evaluated?
4. Define and explain following terms - Precision & Recall.
5. Explain the trade-off between precision and recall.