3 Evaluation and Visualization of Information Retrieval System

UNIT - III

Syllabus

Performance evaluation : Precision and recall, MRR, F-Score, NDCG, user-oriented measures. Visualization in Information System: Starting points, Query Specification, document context, User relevance judgment, Interface support for search process.

3.1 Performance Evaluation

- Information retrieval system is used for searching the relevant documents with respect to the query. It includes following steps :
  - Finding internal representation of documents.
  - Deciding the file structure using which the documents are saved inside the system.
  - Selecting model for matching the documents with the query.
  - Finding internal representation of query.
  - Evaluation of the system.
- Once all the decisions are taken regarding the presentation, organization, and search, the final step before implementation of the IR system is an evaluation of the system.
- The type of evaluation depends on the objective of retrieval system. Thus, the first type of evaluation is **functional analysis** and the second is **error analysis**. Both of these are part of testing phase. Various measures are used for evaluating performance of the system.
- The most common measures of system performance are :

1. Response Time
2. Space

- The shorter the response time, the smaller the space used, the better the system is considered to be.
- For a system designed for providing data retrieval, two basic metrics are used :
  1. **Response time :** The time between a query is submitted to the system and the answer given by the system.
  2. **Space :** Space required to store documents and handle query.
- The factors which affects above measures are :
  - Performance of the indexing structures.

- Interaction with the operating system.
- Delay in communication channels.
- Overhead introduced by many software layers which are usually present.

- Such form of evaluation can be referred as **performance evaluation**.
- In information retrieval system, other metrics, besides time and space, are also used. In IR system, query or the request given the user can be exact or vague. Hence retrieved documents are not the exact answer and need to be ranked according to the relevance to the query. Thus IR systems require the evaluation, which indicates how precise the answer set is. This type of evaluation is referred as **retrieval performance evaluation**.

3.1.1 Retrieval Performance Evaluation

- Here we discuss about the retrieval performance evaluation for information retrieval systems. Such an evaluation is usually based on a test reference collection and evaluation measure.
- The ideal test reference collection consists of :
  1. Collection of documents.
  2. A set of example information requests.
  3. A set of relevant documents (provided by specialists) for each example information requests.
- Given a retrieval strategy S, the evaluation measure quantifies (for each example information request) the **similarity** between the set of documents retrieved by S and the set of relevant documents provided by the specialists. This provides an estimation of the **goodness** of the retrieval strategy S.
- Before considering the evaluation, we must consider which type of task the system will handle :
- The task can be :
  1. Query processed in **batch mode** (i.e. user submits the query and gets the result back).
  2. Query is handled in **interactive session**. Here user will submit the query and see what is in the result. Based on the result given by system, user may refine the query.
  3. Combination of batch and interactive mode.
- In an interactive session, different things may affect the performance of the system and show impact on evaluation. The things which may affect the evaluation are :
  1. User effort
  2. Characteristics of the interface design.
  3. Guidance provided by the system.
  4. Duration of the session.

3.2 Precision and Recall

University Questions

Q. Explain the terms precision and recall. **SPPU : May 15, May 16, 2/5 Marks**

Q. Define and explain following terms : Precision and recall. **SPPU : March 19, 5 Marks**

Recall and precision are the basic measures for retrieval performance evaluation.

Consider

I : Information request (of a test reference collection)

R : Set of relevant documents

|R| : Number of documents in the set of relevant documents.

A : Answer set generated by the information retrieval system.

|A| : Number of documents in the answer set.

|Ra| : Number of documents in the intersection of the sets R and A.

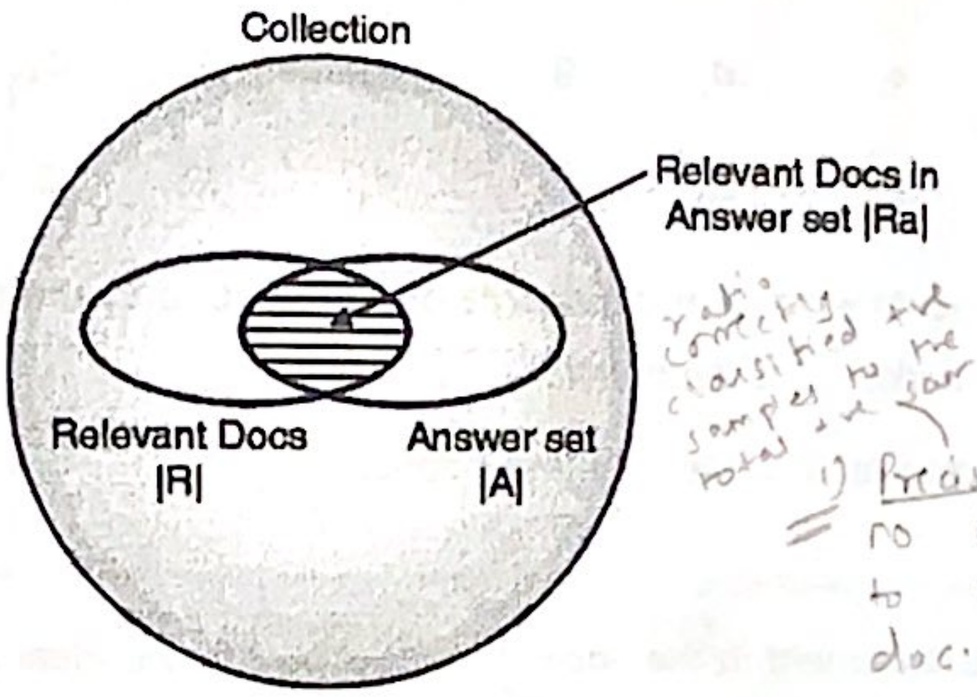


Fig. 3.2.1 : Precision and Recall

Precision and recall for a given example information request

The recall and precision measures are defined as follows :

1. **Recall :** It is the fraction of the relevant documents (the set R) which has been retrieved i.e.

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

2. **Precision :** It is the fraction of the retrieved documents (the set A) which is relevant i.e.

$$\text{Precision} = \frac{|Ra|}{|A|}$$

- The above measures are defined with the assumption that user will go through all the documents in the answer set. But practically, this happens rarely. In real situations, the documents in the answer set are ranked and then displayed to the user. The user examines the documents from top to bottom.
- In such situations recall and precision vary as the user proceeds with his examination of answer set A. Thus, we require plotting the precision vs recall curve.
- Consider following example :

Consider a reference collection and its set of example information request.

q : The query based on the information request.

R_q : Set of relevant documents to the query q, suggested by the specialist.

Example

$$R_q = \{d_1, d_4, d_{15}, d_{39}, d_{44}, d_{46}, d_{72}, d_{88}, d_{92}, d_{110}\}$$

- Consider the new retrieval system. Assume that the algorithm returns the list of documents relevant to the query q. The ranking of the documents in the answer set as follows :
- Ranking for given query :

1.	d_{39}	6.	d_{110}	11.	d_{50}
2.	d_{22}	7.	d_{95}	12.	d_{60}
3.	d_{88}	8.	d_{89}	13.	d_{77}
4.	d_2	9.	d_{112}	14.	d_{45}
5.	d_3	10.	d_4	15.	d_1

- The list represents the documents which is the outcome of search done by the retrieval system. The actual relevant documents to the query are marked unit the bullets.

Let us calculate the precision and recall value for these points.

1. **Document d_{39}**
   - It is ranked as number 1 document in the answer set. It is part of the ideal answer set. Hence we can calculate precision at this document as,
   - There is single document in the answer set out of which the document itself is actually relevant to the query. Hence precision value is 100% (out of one document one document is relevant)
   - Recall can be calculated as the documents which are searched till the first ranked documents are having 10% of actual relevant docs. Hence recall is 10%.
2. **Document d_{88}**
   - Now at these points we are having 3 documents in the answer set i.e. d_{39}, d_{22} and d_{88}. Out of these 3 documents (d_{39} and d_{88}) 2 documents are actually relevant. Hence precision is 66.66% (2 out of 3).
   - Recall can be calculated as, till yet we found 2 documents which are part of ideal answer set. The ideal answer set contains 10 documents. Hence recall is 20% (2 out of 10). Thus, the precision and recall values for the list are :

Document		Precession	Recall
(d_{39})	(1 out of 1)	100%	10%
(d_{88})	(2 out of 3)	66.66%	20%
(d_{110})	(3 out of 6)	50%	30%
(d_4)	(4 out of 10)	40%	40%
(d_1)	(5 out of 15)	33.33%	50%

- We can draw a graph based on the above values.
- Note that precision values drops to 0% for recall value above 50% because further there is no relevant doc. present in the list generated by system.
- In the above example, the precision and recall figures are for a single query. Usually, 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.

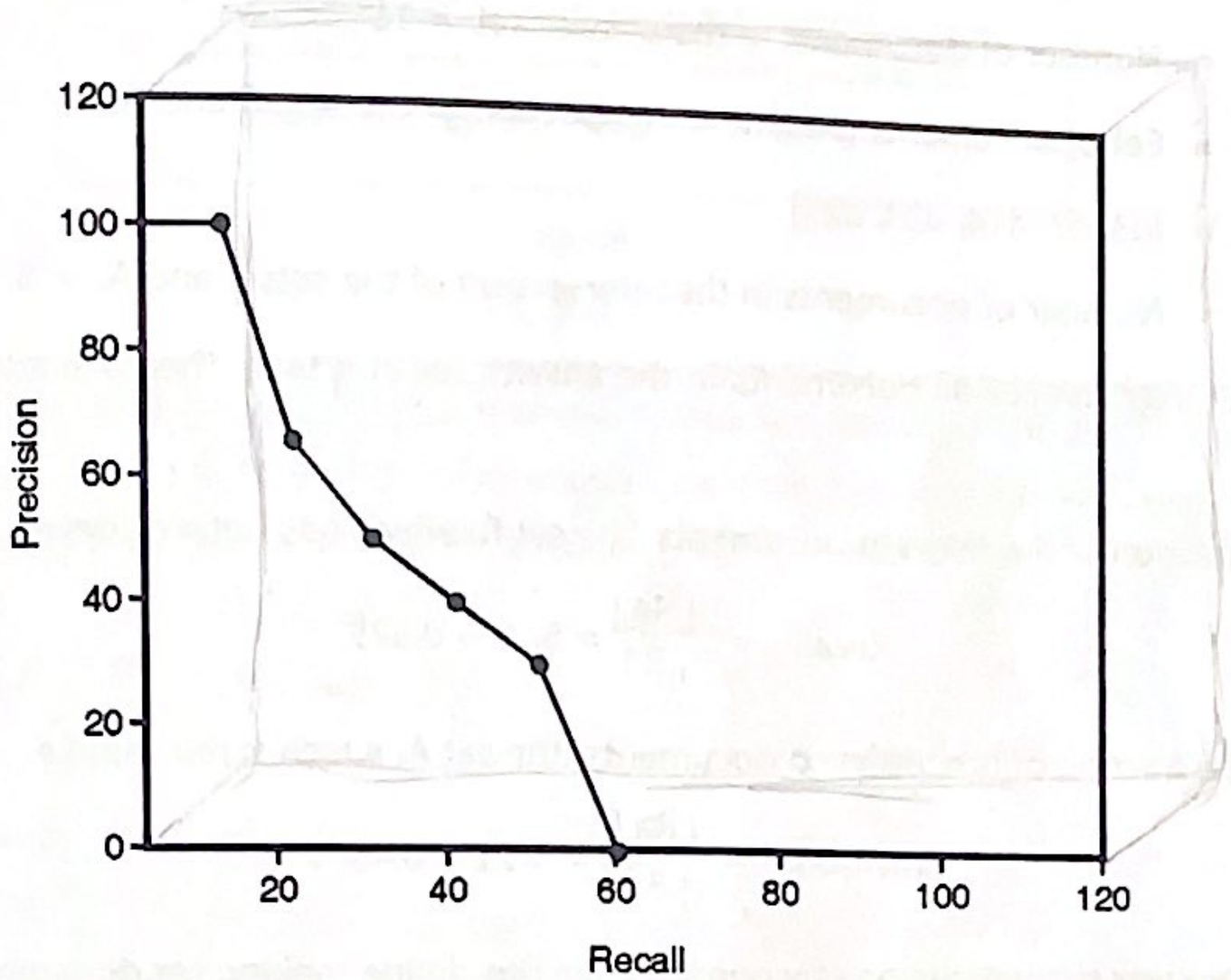


Fig. 3.2.2

- Note that precision values drops to 0% for recall value above 50% because further there is no relevant doc. present in the list generated by system.
- In the above example, the precision and recall figures are for a single query. Usually, 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 can average the precision figures at each recall level as follows :

$$\bar{P}(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q}$$

Where, $\bar{P}(r)$: Average precision at recall level r

N_q : Number of queries used.

$P_i(r)$: Precision at recall level r for i^{th} query

Ex. 3.2.1 : Explain the term precision and recall and calculate the same for following example. The set of relevant documents in the query Q = {d3, d7, d8, d11, d14, d19, d23, d25}
A new retrieval algorithm returns following answer set = {d1, d2, d3, d7, d9, d10, d14, d20, d23, d24, d25}

SPPU - May 15, May 16, 5 Marks

Soln. :

Let R = Set of relevant documents = {d3, d7, d8, d11, d14, d19, d23, d25}

|R| = Number of documents in the set of relevant documents = 8

A = Answer set of algorithm

= {d1, d2, d3, d7, d9, d10, d14, d20, d23, d24, d25}

|A| = Number of documents in the answer set = 11

Ra = Set of documents present in intersection of the sets R and A.

= {d3, d7, d14, d23, d25}

|Ra| = Number of documents in the intersection of the sets R and A. = 5

- If we assume that user can see all documents in the answer set at a time, then a single value will be for precisior and recall.

1. **Recall :** It is the fraction of the relevant documents (the set R) which has been retrieved i.e.

$$\text{Recall} = \frac{|Ra|}{|R|} = 5/8 = 0.625$$

2. **Precision :** It is the fraction of the retrieved documents (the set A) which is relevant i.e.

$$\text{Precision} = \frac{|Ra|}{|A|} = 5/11 = 0.4545$$

But if our algorithm lists documents one by one then we can define ranking for documents based on the sequence in which documents are listed: Thus the ranking of documents will be :

1.	d_1	6.	d_{10}	11.	d_{25}
2.	d_2	7.	d_{14}		
3.	d_3	8.	d_{20}		
4.	d_7	9.	d_{23}		
5.	d_9	10.	d_{24}		

Document	Precision	Precision	Recall	Recall
(d_3)	(1 out of 3)	33.33%	(1 out of 8)	12.5%
(d_7)	(2 out of 4)	50%	(2 out of 8)	25%
(d_{14})	(3 out of 7)	42.85%	(3 out of 8)	37.5%
(d_{23})	(4 out of 9)	44.44%	(4 out of 8)	50%
(d_{25})	(5 out of 11)	45.45%	(5 out of 8)	62.5%

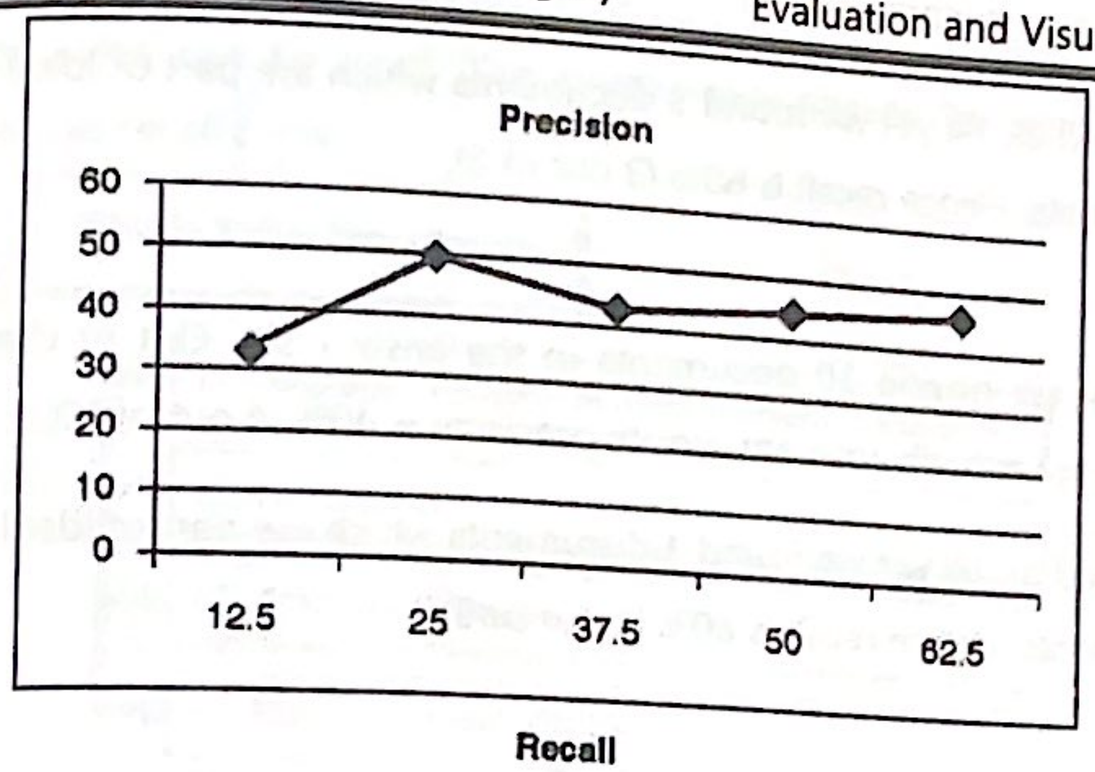


Fig. P.3.2.1

Ex. 3.2.2 : Consider a reference collection and its set of example information request. If q is the information request and a set Rq = (d3, d5, d9, d25, d39, d44, d50, d70, d80, dl20). Now consider new retrieval algorithm has been designed and has been evaluated for information request q returns, ranking of the documents in the answer set as.

1. d120	6. d9	11. d38
2. d84	7. d58	12. d48
3. d50	8. d129	13. d230
4. d6	9. d143	14. d113
5. d8	10. d25	15. d3

The documents that are relevant to the query q are underlined. Calculate precision and recall for the documents that are relevant to the query q.

SPPU : Dec. 16, May 17, 10 Marks

Soln. :

To calculate the precision and recall value for the documents that are relevant to the query q, following steps need to be follow:

1. **Document d_{120}**
   - It is ranked as number 1 document in the answer set. It is part of the ideal answer set. Hence we can calculate precision at this document as,
   - There is single document in the answer set out of which the document itself is actually relevant to the query. Hence precision value is 100% (out of one document one document is relevant)
   - Recall can be calculated as the documents which are searched till the first ranked documents are having 20% of actual relevant docs. Hence recall is 20%.
2. **Document d_{50}**
   - Now at this point we are having 3 documents in the answer set i.e. d_{120}, d_{84} and d_{50}. Out of these 3 documents (d_{120} and d_{50}) 2 documents are actually relevant. Hence precision is 66.66% (2 out of 3).
   - Recall can be calculated as, till yet we found 2 documents which are part of ideal answer set. The ideal answer set contains 5 documents. Hence recall is 40% (2 out of 5).
3. **Document d_9**
   - Now at these points we are having 6 documents in the answer set i.e. d_{120}, d_{84}, d_{50}, d_6, d_8 and d_9. Out of these 6 documents (d_{120}, d_{50}, and d_9) 3 documents are actually relevant. Hence precision is 50% (3 out of 6).

- Recall can be calculated as, till yet we found 3 documents which are part of ideal answer set. The ideal answer set contains 5 documents. Hence recall is 60% (3 out of 5).

4. Document d_{25}

- Now at this point we are having 10 documents in the answer set. Out of these 10 documents (d_{12}, d_{50}, d_9, and d_{25}) 4 documents are actually relevant. Hence precision is 40% (4 out of 10).
- Recall can be calculated as, till yet we found 4 documents which are part of ideal answer set. The ideal answer set contains 5 documents. Hence recall is 80% (4 out of 5).

5. Document d_3

- Now at this points we are having 15 documents in the answer set. Out of these 15 documents (d_{12}, d_{50}, d_9, d_{25} and d_3) 5 documents are actually relevant. Hence precision is 33.33% (5 out of 15).
- Recall can be calculated as, up till we found 5 documents which are part of ideal answer set. The ideal answer set also contains 5 documents. Hence recall is 100% (5 out of 5).
- Thus, the precision and recall values for the list are :

Document		Precession	Recall
(d_{120})	(1 out of 1)	100%	20%
(d_{50})	(2 out of 3)	66.66%	40%
(d_9)	(3 out of 6)	50%	60%
(d_{25})	(4 out of 10)	40%	80%
(d_3)	(5 out of 15)	33.33%	100%

3.3 Single Value Summaries

University Question

Q. Explain the term

(i) R precision

(ii) Precision Histogram

SPPU : Dec. 14, May 16, May 19, 3/4 Marks

- Average precision versus recall figures are useful for comparing the retrieval performance of distinct retrieval algorithms.
- However there are situations in which we would like to compare the retrieval performance of our retrieval algorithms for the individual queries.
- Reasons are :

1. Averaging precision over many queries might disguise important anomalies in the retrieval algorithms under study.
2. When we want to compare two algorithms, investigating whether one of them out performs the other for each query in a given set of example queries.

- In these situations, single value can be used. This single value should be interpreted as a summary of the corresponding precision versus recall curve.

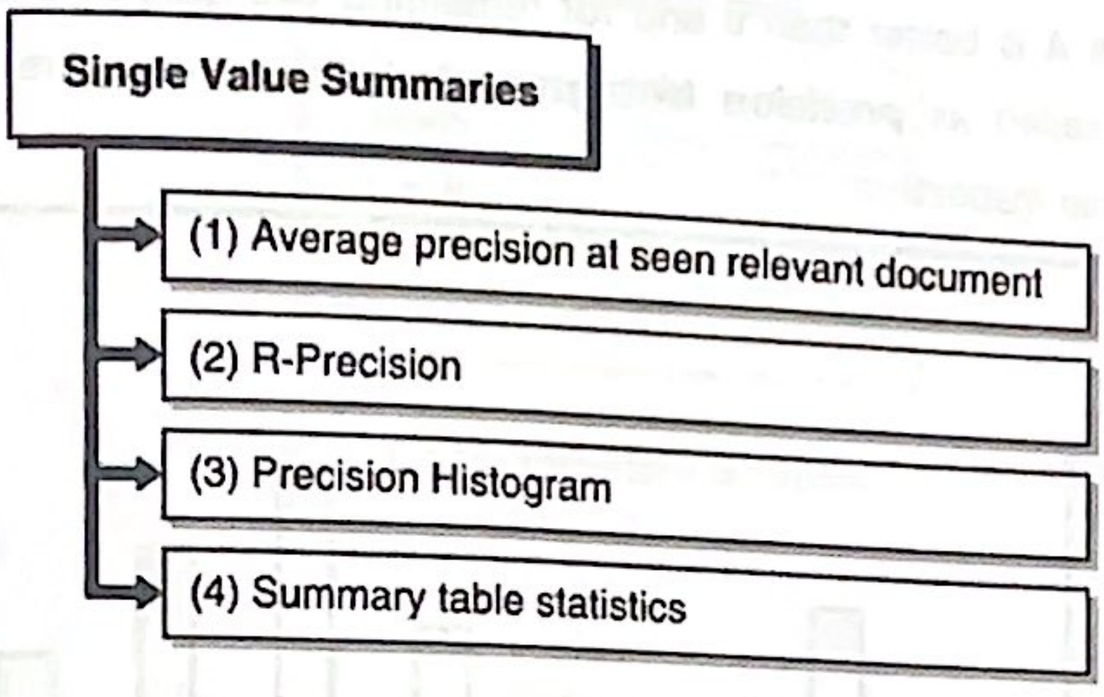


Fig. 3.3.1 : Single value summaries

1. Average precision at seen relevant document

- It is the average of precision figures obtained after each new relevant document is observed.
- For example in above example the precision figures after each new relevant document are 1, 0.66, 0.5, 0.4 and 0.33. Thus, **the average precision of seen documents** is given by,

$$\frac{1 + 0.66 + 0.5 + 0.4 + 0.33}{5} = 0.578$$

- This measure favours systems which retrieve relevant documents quickly (i.e. early in the ranking.) An algorithm might present good precision but may have poor performance in terms of recall.

2. R-Precision

- R-precision is the precision value at R^{th} level, where R is the total number of relevant documents for the current query (i.e. number of documents in the set R_q).
- Consider the above example. The value of R-precision is 0.4 for this example (because R = 10 and there are four relevant documents out of first 10 documents). The R-precision measure is useful for checking the behaviour for single query.

3. Precision Histogram

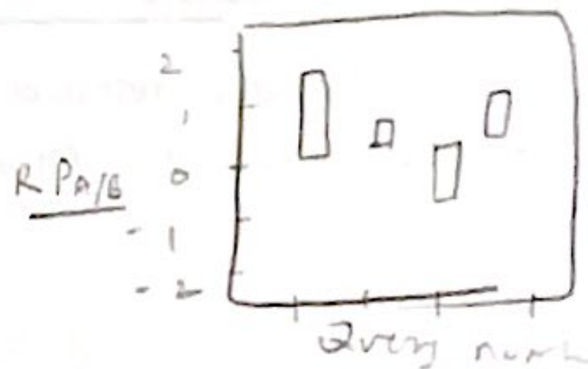
- The R-precision values for several queries can be used to compare the performance of two different algorithms.
- Let, we have two algorithms A and B.

$RP_A(i)$ = R-precision value of algorithm A for i^{th} query.

$RP_B(i)$ = R-precision value of algorithm B for the i^{th} query.

Then calculate, $\mathbf{RP_{A/B}(i) = RP_A(i) - RP_B(i)}$

- Let consider three different outcomes :
  1. $RP_{A/B}(i) = 0$, indicates both algorithms have equivalent performance.
  2. $RP_{A/B}(i) > 0$, indicates better retrieval performance by A compared with B.



3. $RP_{A/B}(i) < 0$, indicates better retrieval performance by B, than A.

- For 8 queries algorithm A is better than B and for remaining two queries B algorithm is better than A. This type of bar graph is called as **precision histogram**. It helps to compare performance of two different algorithms through visual inspection.

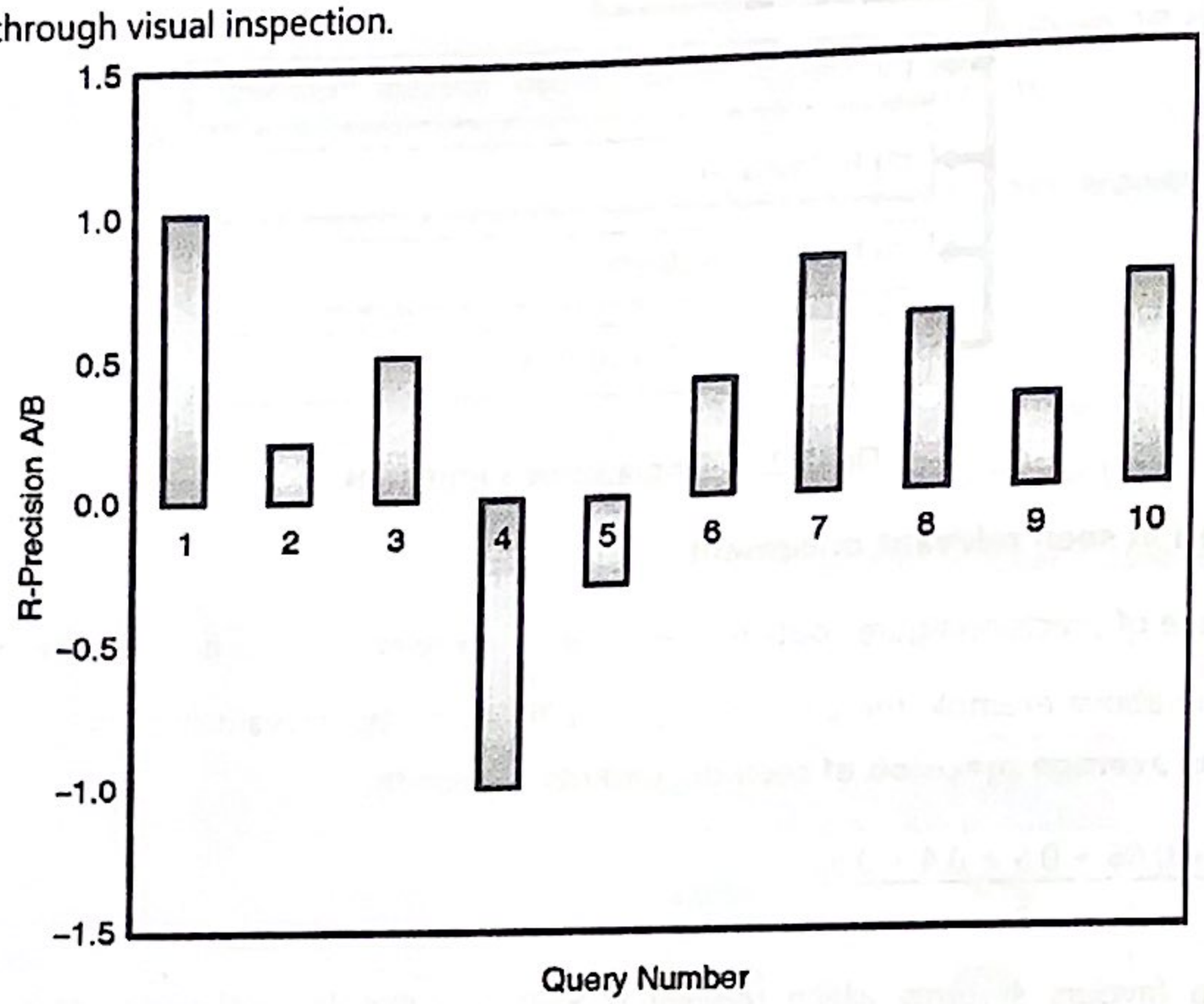


Fig. 3.3.2

4. Summary table statistics

Single value summaries can be stored in a table to provide statistical information about the performance. This table may includes the information like :

1. Number of queries used in the task.
2. Total number of documents.
3. Total number of documents retrieved by all queries.
4. Total number of relevant documents which were effectively retrieved when all queries are considered
5. Total number of relevant documents which could have been retrieved by all queries, etc.

3.4 Alternative Measures

Q. Explain the terms :

(i) Harmonic Mean

(ii) E measure

SPPU : Dec. 14, May 16, May 19, 3/4 Marks

For information retrieval system, precision and recall are basic measures. But these are always not the appropriate measures for evaluation. Other measures which are used are as follows :

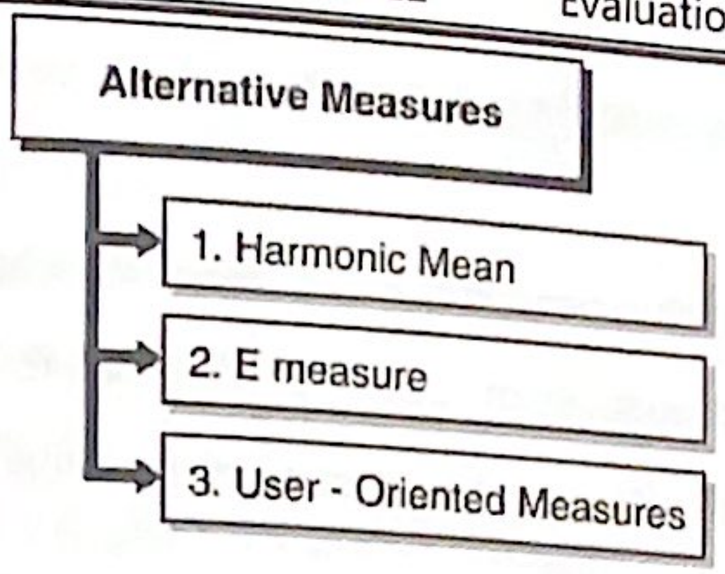


Fig. 3.4.1 : Alternative measures

3.4.1 Harmonic Mean

or F-measure

Let, j = j^{th} document in the collection.

$r(j)$ = Recall for j^{th} document in the ranking.

$P(j)$ = Precision for j^{th} document in the ranking.

$F(j)$ = harmonic mean of $r(j)$ and $p(j)$.

$$\therefore \quad F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}}$$

$$\frac{2}{\frac{1}{P(j)} + \frac{1}{R(j)}}$$

The function F(j) will have the value in the limit from 0 to 1. It is 0 when no relevant documents have been retrieved and is 1 when all ranked documents are relevant. F(j) will have high value when both precision and recall have high value.

3.4.2 E Measure

University Question

Q. Define and explain following terms : i. MRR ii. NDCG. **SPPU : March 19, 5 Marks**

This is another measure to specify the performance of information retrieval system. Here, user may specify whether he is more interested in precision or recall or both.

Let, j = j^{th} document in the collection

$r(j)$ = recall for the j^{th} document in the ranking

$P(j)$ = precision for j^{th} document in the ranking.

$E(j)$ = E-measure relative to $r(j)$ and $P(j)$

b = User specified parameter which shows the relative importance of recall and precision.

$$E(j) = 1 - \frac{1 + b^2}{\frac{b^2}{r(j)} + \frac{1}{P(j)}}$$

b = 0

b → ∞

For b = 1, E(j) is reverse of Harmonic mean F(j).

b > 1, User is more interested in precision then recall.

b < 1, User is more interested in recall than precision.

A. Mean Reciprocal Rank (MRR)

- It is statistical parameter.
- It is used to evaluate any process that generates list of results for a user query.
- This list should be arranged by probability of success or relevance to the user query.
- It is inverse in multiplicative form related to first answer of the query.
- Equation of MRR is,

$$MRR = 1/|n| \left(\sum_{i=0}^{n} 1/rank(i) \right)$$

Where,

n is number of user queries being processed.

rank (i) is position of most relevant document in i^{th} query.

B. F-Score

- This is one of the most important evaluation parameter
- It is based on both precision and recall
- So, it shows type I and type II errors
- F1 score = (precision * recall) *2 / (precision + recall)

C. Normalized Discounted Cumulative Gain (NDCG)

- User queries can be multiple and from multiple domains.
- So, each query may result in variegated number of matches.
- Thus evaluating results may require different measures than which are used traditionally.
- Gain at each position should be normalized.
- First step is, documents are sorted by their relevance.
- Cumulative gain at each position is calculated.
- Then this value is normalized for each position.
- This measure is very useful because it will penalize those relevant documents which appear low in rank.
- Such penalization encourages better relevance based ranking.
- Problem with this approach is it does not penalize bad documents.

3.4.3 User - Oriented Measures

University Questions

Q. Define and explain following concepts : User oriented measures for performance evaluation. **SPPU : March 19, 3 Marks**

Q. Write a short note on user oriented measures to evaluate the performance of the system. **SPPU : May 17, 4 Marks**

- The values of precision and recall are calculated considering that set of relevant document for a query is the same, independent of the user. But practically speaking, the decision of relevant documents may vary user to user.

- Hence some other measures are defined which consider the user's point of view.
- Let consider a reference collection.

I = Example information request

R = Set of relevant documents for I.

A = Answer set retrieved.

U = U be the subset of R which is known to the user.

|U| = Number of documents in set U.

R_k = A ∩ U

= Documents known to the user to be relevant which were retrieved.

$|R_k|$ = Number of documents in the set R_k.

R_u = Documents previously unknown to the user which were retrieved.

$|R_u|$ = Number of documents in the set R_u.

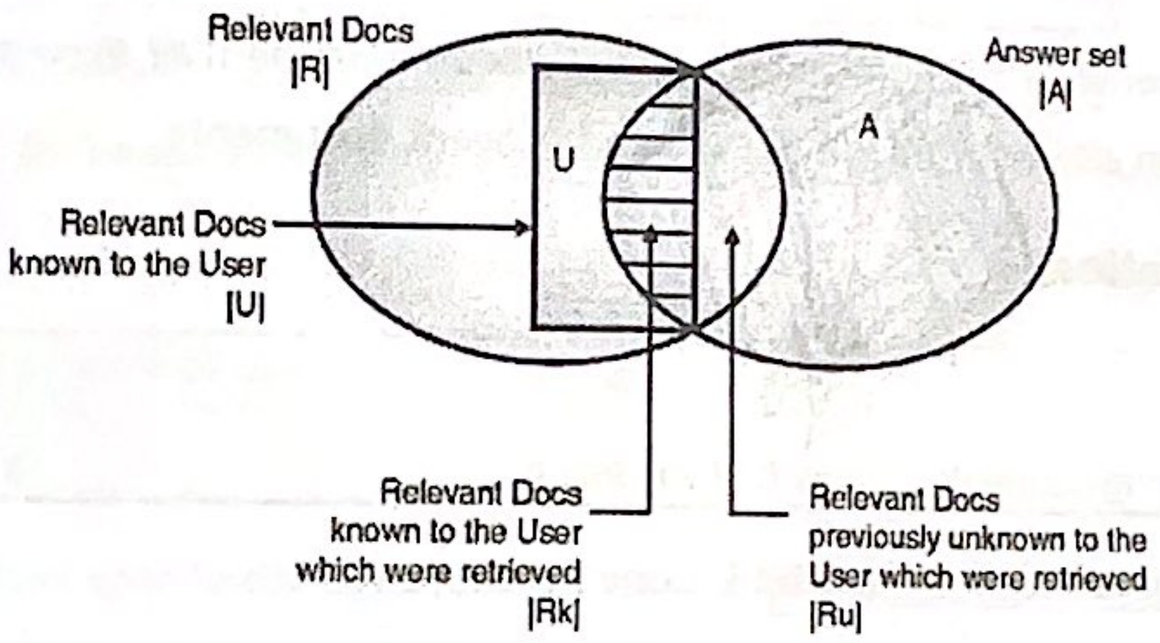


Fig. 3.4.2

Different measures

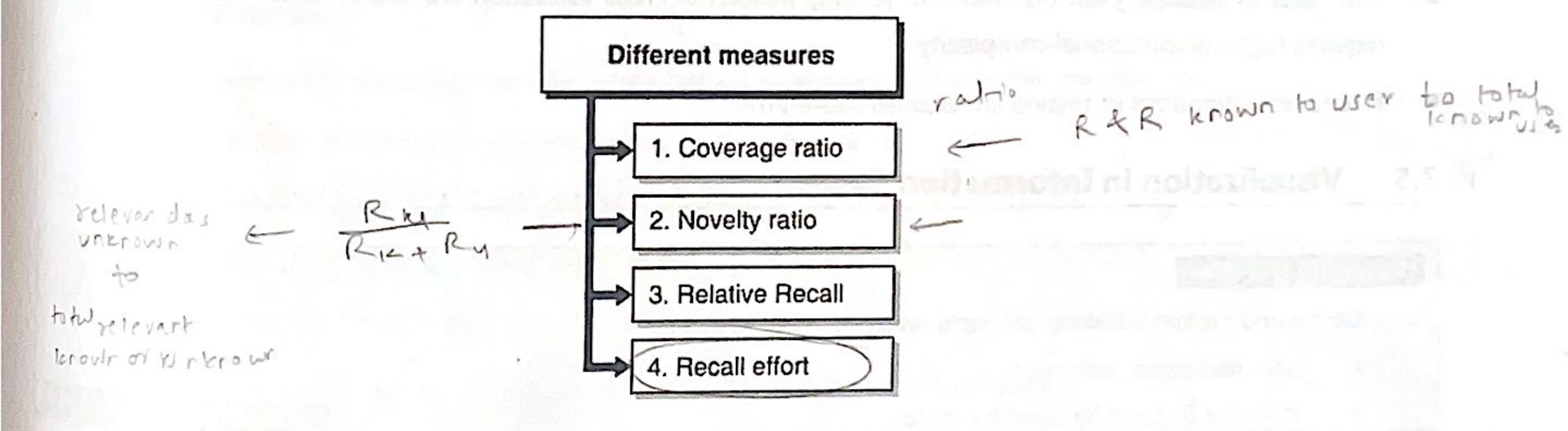


Fig. 3.4.3 : Different measures

1. Coverage ratio

Coverage ratio can be defined as the fraction of the documents known (to the user) to be relevant which has actually been retrieved i.e.

$$\text{Coverage} = \frac{|R_k|}{|U|}$$

2. Novelty ratio

- Novelty ration is defined as the fraction of the relevant documents retrieved which was unknown to the user i.e.

$$\text{Novelty} = \frac{|R_u|}{|R_u| + |R_k|}$$

- A high coverage ratio indicates that system is finding most of the relevant document the user expected to see.

  A high novelty ratio indicates that the system is finding many new relevant documents which were previously unknown to the user.

3. Relative Recall

Relative recall is the ratio between the number of relevant documents found by the system and the number of relevant document user is expected to find. When user finds as many document as he expected, he will stop search and relative value will be 1.

4. Recall effort

Recall effort is the ratio between the number of relevant documents the user expected to find and the number of documents examined in an attempt to find the expected relevant documents.

3.4.4 Cross Fold Evaluation

University Question

Q. Define and explain following concepts : Cross fold validation. **SPPU : March 19, 2 Marks**

- Evaluation using various parts (folds) of dataset is done here.
- First dataset is split in 'k' parts. Each of these k folds is used for testing. It takes total k iterations for complete testing
- This helps in averaging out the results of testing. Results of cross validation are useful bench mark of evaluation. It requires high computational complexity.
- It requires k iterations of testing so requires more time.

3.5 Visualization in Information System

University Question

Q. Define and explain following concepts related to visualization.

i. User relevance Judgment

ii. Interface Support for search process **SPPU : March 19, 5 Marks**

- Visualization is important for understanding the data. Various aspects of documents can be visualized.
- User relevance and interface support can also be one aspect. Starting points of visualization can be a key for visualization.

1. Starting Points

- Information visualization is important to find patterns in data
- It serves as tool to lead to decisions in various aspects.

2. Document Context

- Document context can be visualized in various forms
- Term frequency is one of the way to visualize the context
- Another common way is using binary representation
- Using these representation documents can be visualized

3. User Relevance Judgment

- Relevance of document with user query is very good measure for evaluation of document retrieval system
- It is measured using following factors :
  a. Time spent with the document
  b. Click on the link or not
  c. Explicit survey of relevance
  d. Feedback of relevance from user
- This measure is used to measure performance of any retrieval system.
- This measure can be used to visualize results of retrieval system.
- There can be implicit or explicit user relevance judgment done.

i. Explicit Relevance

- Here user is asked specifically about the relevance of the retrieved results.
- Many times user may not have time or motivation to reply.
- But here user is well informed about relevance feedback.

ii. Implicit Relevance

- Here without informing the user, some behavioural aspects of the user are collected.
- These aspects can be time spent on link or clicking of a link.
- This leads to best relevance feedback.
- But here user is not informed about gathering of parameters.

4. Interface Support for Search Process

- Search process gives results of documents which are relevant to the user query.
- These results need to be presented properly to the user.
- Results can be sorted by order of relevance.
- Another parameter can be frequently used documents.
- Interface should allow tracing of user relevance parameters.
- Interface should allow tracking of implicit as well as explicit user relevance parameters.