IR UNIT – 6 (Measuring Effectiveness and Measuring Efficiency) – END-SEM PYQ Answers

► **NOV/DEC 2022**

Q5) a) Explain Traditional effectiveness measure and The Text Retrieval Conference (TREC) with suitable examples. [9]

Information Retrieval (IR) systems are evaluated by measuring how effectively they retrieve relevant documents. Traditional measures focus on relevance, completeness, and accuracy of retrieval.

1. Traditional Effectiveness Measures

i) ***Precision***

Precision = proportion of retrieved documents that are relevant.

$$Precision = \frac{\text{Relevant Retrieved}}{\text{Total Retrieved}}$$

Example: System retrieves 10 docs; 7 are relevant → Precision = 7/10 = 0.7

ii) ***Recall***

Recall = proportion of relevant documents retrieved by the system.

$$Recall = \frac{\text{Relevant Retrieved}}{\text{Total Relevant}}$$

Example: Total relevant docs = 20; system retrieves 7 → Recall = 7/20 = 0.35

iii) ***F-measure***

Combines precision and recall.

$$F = \frac{2\,PR}{P + R}$$

Balances both measures.

iv) ***Precision–Recall Curve***

Graph showing precision at different recall levels.
Used to compare retrieval methods.

v) ***Average Precision (AP)***

Average of precision values at points where relevant documents appear.

2. The Text Retrieval Conference (TREC)

TREC is a benchmark evaluation initiative established by NIST (National Institute of Standards and Technology).
Goal: Provide a standard platform for evaluating large-scale IR systems.

Features of TREC

1. **Large test collections:** e.g., newswire, web crawl datasets
2. **Standard queries (topics):** Example topics: “Airline accidents”, “Global warming”, etc.
3. **Relevance judgments (qrels):** Documents are labeled as relevant or non-relevant by human judges.
4. **Tracks (specialized tasks)**
    - Web Track
    - Question Answering Track
    - Spam Track
    - Medical Track

Example of a TREC Task

Topic: *“Impact of climate change on agriculture”*
System retrieves top 1000 documents → TREC compares system’s ranking with gold-standard relevance judgments using measures like MAP, NDCG, Recall.

Importance

- Industry standard for comparing IR systems
- Provides large-scale, unbiased evaluation
- Encourages research innovation

b) Write a short note on: [9]
i) Nontraditional effectiveness measures **ii) Measuring efficiency**

i) Nontraditional Effectiveness Measures: These measures evaluate IR systems beyond classical precision/recall, considering *user behavior, interaction, rank quality, and satisfaction*.

Types

1. **Rank-based Metrics**

- NDCG (Normalized Discounted Cumulative Gain):
  Rewards relevant documents at top ranks.
- MRR (Mean Reciprocal Rank):
  Used for question answering; measures rank of first correct answer.

2. **User-Satisfaction Measures**

- Click-through rate (CTR)
- Session success rate

- Dwell time on pages

3. **Utility-based Measures:** Evaluate usefulness of results rather than strict relevance.

4. **Diversity and Novelty Metrics**

- Ensuring diverse topics in search results.
- Avoiding duplicates.

Use Cases: Modern search engines (Google, Bing), recommender systems, web search ranking.

ii) Measuring Efficiency: Efficiency measures focus on speed, resource usage, and scalability rather than relevance.

Key Aspects

1. **Query Response Time**: Time to process a query.
2. **Indexing Time**: Time to build or update index.
3. **Memory Usage**: Space required to store index structures.
4. **Throughput**: Number of queries processed per second.
5. **CPU Cost & I/O Cost**: Determines ability to handle large data.

Example: Engine A returns results in 50 ms; Engine B takes 200 ms → A is more efficient.
Efficiency ensures real-time responsiveness of IR systems in real-world applications.

Q6) a) What is Scheduling and Caching in Measuring Efficiency?. Explain in detail. [9]

Measuring efficiency includes optimizing system resources to handle large-scale queries. Two major techniques are scheduling and caching.

1. Scheduling: Scheduling determines how queries or indexing jobs are ordered and executed to minimize latency and maximize throughput.

Scheduling Techniques

a) ***FIFO (First In, First Out)***

Queries processed in arrival order.

b) ***Priority Scheduling***

- Certain queries get higher priority (e.g., user-facing queries over batch indexing).
- Critical tasks handled first.

c) ***Round-Robin Scheduling:*** Used when multiple processors or servers are involved.

d) ***Load Balancing:*** Queries distributed evenly across servers to avoid bottlenecks.

Benefits

- Reduces response time
- Handles peak loads
- Improves system throughput

2. Caching: Caching stores frequently accessed data in fast memory to reduce retrieval cost.

Types of Caching in IR Systems

a) ***Query Result Cache:*** Stores results of frequent queries (e.g., “weather”, “cricket score”).

b) ***Inverted List Cache:*** Frequently accessed posting lists stored in memory.

c) ***Document Cache:*** Stores popular or recently visited documents.

d) ***Disk Cache:*** Reduces disk read operations.

Cache Replacement Policies

- LRU (Least Recently Used)
- LFU (Least Frequently Used)
- MRU (Most Recently Used)

Benefits

- Faster response time
- Less I/O overhead
- Reduces CPU and disk usage
- Essential for large-scale IR systems like Google search

b) Write a short note on: [9]
i) Using statistics in evaluation **ii) Minimizing adjudication Effort**

i) Using Statistics in Evaluation: Statistics help determine the significance and reliability of IR system performance.

Uses of Statistics in IR Evaluation

1. **Confidence Intervals:** Show reliability of performance metrics (e.g., MAP ± standard error).
2. **Statistical Significance Tests**
    - Paired t-test
    - Wilcoxon signed-rank test

Used to check if improvement over baseline is real or due to randomness.

3. **Randomization Tests:** Useful for comparing ranked retrieval outputs.
4. **Variance and Standard Deviation:** Measure stability of system performance across many queries.

Outcome: Ensures fair comparison of IR systems.

ii) Minimizing Adjudication Effort

Adjudication = human effort needed to judge whether documents are relevant.

Challenges

- Very expensive
- Time-consuming
- Difficult for large datasets such as web crawls

Techniques to Minimize Effort

1. **Pooling Method (Used in TREC):** Only top-ranked results from multiple systems are judged → reduces total judgment load.

2. **Active Learning:** System selects the most informative documents for human assessment.

3. **Sampling:** Judge only a small sample of documents.

4. **Reusing Judgments:** Same judgment used for multiple evaluations if topics overlap.

5. **Automatic (Machine) Judging:** Using ML models to predict relevance where human judgment is too costly.

Benefits

- Reduces human workload
- Improves evaluation feasibility
- Saves cost and time
- Enables large-scale IR experiments

► **MAY/JUNE 2023**

Q5) a) Explain Measuring Effectiveness like Traditional effectiveness measure and the text retrieval conference (TREC) with suitable examples. [9] → DONE

b) Write a short note on: [9]
i) Nontraditional effectiveness measures → DONE **ii) Measuring efficiency → DONE**

Q6) a) Explain Query Scheduling and Caching with suitable examples. [9] → DONE

b) Write a short note on: [9]
i) Using statistics in evaluation → DONE **ii) Minimizing adjudication Effort → DONE**

- **NOV/DEC 2023**

Q5) a) Explain Traditional effectiveness measure and The Text Retrieval Conference (TREC) with suitable examples. [9] → DONE

b) Write a short note on: [9]
i) Nontraditional effectiveness measures → DONE **ii) Measuring efficiency → DONE**

Q6) a) What is Scheduling and Caching in Measuring Efficiency?. Explain in detail. [9] → DONE

b) Write a short note on: [9]
i) Using statistics in evaluation → DONE **ii) Minimizing adjudication Effort → DONE**

- **MAY/JUNE 2024**

Q5) a) Explain Traditional effectiveness measure and The Text Retrieval Conference (TREC) with suitable examples. [9] → DONE

b) Write a short note on: [9]
i) Redis and Memcached **ii) Measuring efficiency**

(i) Redis and Memcached [~4-5 marks]

Redis and Memcached are two widely used in-memory caching systems that significantly improve the performance and efficiency of information retrieval (IR) systems, web applications, and large-scale databases.

1. Redis

Redis (REmote DIctionary Server) is an in-memory key-value store that supports advanced data structures.

Key Features

- Stores data in forms like strings, lists, sets, sorted sets, hashes, bitmaps, etc.
- Very fast reads/writes → microsecond latency.
- Supports persistence (RDB snapshots, AOF logs).
- Includes built-in replication and high availability (Redis Cluster).

Use Cases

- Query caching in search engines
- Storing session data

- Caching inverted lists
- Leaderboards (sorted sets)

Example: If a query like “Olympics 2024 results” is frequently asked, Redis can store the top results so the system responds instantly without re-running ranking algorithms.

2. Memcached

Memcached is a high-performance distributed memory object cache used primarily for simple caching.

Key Features

- Simple key-value storage
- Extremely fast for storing small objects
- No persistence (cache-only)
- Scales horizontally very easily

Use Cases

- Database query caching
- Web page caching
- Temporary storage of popular search results

Example: A website stores user profile data (after first database fetch) in Memcached → all next requests load data directly from memory.

Summary

Feature	**Redis**	**Memcached**
Data Structures	Multiple (list, set, hash)	Key-value only
Persistence	Yes	No
Performance	Very fast	Slightly faster for simple keys
Use Case	Complex caching	Simple caching

Conclusion: Both Redis and Memcached greatly improve search engine response time, reduce load on servers, and increase efficiency by storing frequently accessed data in memory.

(ii) Measuring Efficiency [~4.5 marks]

Measuring efficiency evaluates how fast and resource-effective an IR system is while processing queries, indexing documents, or managing updates. Unlike effectiveness (precision/recall), efficiency focuses on speed, memory, CPU usage, and scalability.

1. **Query Response Time**
   - Time taken to return results for a user query.
   - Lower is better.
     Example: Google responds in <50 ms.
2. **Indexing Time**
   - Time needed to build or update the inverted index.
   - Needs to be fast for dynamic collections (news, web data).
3. **Throughput**
   - Number of queries processed per second.
   - Indicates system capacity during peak load.
4. **Memory and Space Usage**
   - Size of the index after compression.
   - Efficient data structures reduce RAM/disk usage.
5. **CPU and I/O Costs**
   - Measures compute load and disk read/write operations.
   - Lower cost → more scalable system.
6. **Caching & Scheduling**
   - Caching (Redis/Memcached) reduces computation by storing frequent results.
   - Scheduling ensures efficient allocation of system resources.
7. **Latency vs. Scalability**
   - Latency: time for one query
   - Scalability: ability to handle many queries simultaneously

High-efficiency IR systems maintain low latency even under heavy load.

Conclusion: Efficiency metrics ensure that an IR system is not only accurate but also fast, scalable, and resource-optimized, enabling real-time search performance.

Q6) a) What is Scheduling and Caching in Measuring Efficiency? Explain in detail. [9] → DONE

b) Write a short note on: [9]
i) Using statistics in evaluation → DONE **ii) Minimizing adjudication Effort → DONE**

► **NOV/DEC 2024**

Q5) a) Explain Measuring effectiveness like Traditional effectiveness measure and the text retrieval conference (TREC) with suitable examples. [6]

Measuring effectiveness checks **how well** an IR system retrieves relevant documents. Two major methods are:

1) Traditional Effectiveness Measures

i) Precision

Measures accuracy of retrieved results.

$$Precision = \frac{\text{Relevant Retrieved}}{\text{Total Retrieved}}$$

Example: Retrieved 10 docs, 7 relevant → Precision = 0.7

ii) Recall

Measures completeness.

$$Recall = \frac{\text{Relevant Retrieved}}{\text{Total Relevant}}$$

Example: Total relevant = 20, retrieved = 7 → Recall = 0.35

iii) F-Measure

Balances precision and recall.

$$F = \frac{2\,PR}{P + R}$$

Used when both measures are equally important.

iv) Precision–Recall Curve: Shows precision values at different recall levels to compare systems.

2) TREC (Text Retrieval Conference)

TREC is a large-scale IR evaluation initiative by NIST.

Features

- Provides standard **datasets**, **queries**, and **relevance judgments**
- Ensures **fair comparison** between IR systems
- Introduced concept of **evaluation tracks**: Web Track, QA Track, Spam Track, etc.

Example: A system retrieves results for TREC query "World Oil Prices."
Its ranking is evaluated using TREC's official measures like MAP, NDCG, recall.

b) Write a short note on: [6]

i) Nontraditional effectiveness measures ii) Measuring efficiency

i) Nontraditional Effectiveness Measures (3 Marks)

These measures go beyond classical precision/recall and consider **ranking, user behavior, diversity, and usefulness**.

Types

- **NDCG (Normalized Discounted Cumulative Gain):**
  Rewards relevant docs appearing at top ranks.
- **MRR (Mean Reciprocal Rank):**
  Used in QA systems; evaluates rank of *first* relevant answer.
- **Session-Based Measures:**
  Include click-through rate, user dwell time.
- **Diversity Measures:**
  Ensure search results cover multiple subtopics.

Used in: Modern search engines like Google, Bing.

ii) Measuring Efficiency (3 Marks)

Efficiency measures focus on system **speed, scalability, and resource usage**.

Key Metrics

- **Query Response Time** – time taken to answer a query
- **Indexing Time** – time to build/update inverted index
- **Throughput** – number of queries handled per second
- **CPU & I/O Cost** – computational cost
- **Memory Usage** – index storage and RAM usage

Example: System A responds in 50 ms, System B in 200 ms → A is more efficient.

c) Explain query scheduling with suitable examples. [6]

Query scheduling determines **how queries are ordered and executed** to ensure fast response and balanced system load.
1) FIFO Scheduling (First-In, First-Out): Queries processed in arrival order.
Example: A ticket booking system serves users sequentially without priority.

2) Priority-Based Scheduling: Some queries get higher priority.
Example:

- Real-time user queries over background indexing jobs
- Payment gateway queries over informational queries

3) Load-Balanced Scheduling: Queries distributed across multiple servers to avoid overload.
Example: Google routes search requests to the nearest or least busy data center.

4) Round-Robin Scheduling: Equal time slices given to each queue/server. Useful in distributed IR systems.

Benefits

- Faster query response
- Better resource utilization
- Prevents server overload
- Maintains fairness between user tasks and background tasks

Q6) a) Write a short note on: [6]

i) Minimizing adjudication effort ii) Using statistics in evaluation

i) Minimizing Adjudication Effort (3 Marks)

Adjudication = human effort to judge whether documents are relevant.

Challenges

- Time-consuming
- Costly
- Not scalable for large datasets (e.g., web crawls)

Techniques

- **Pooling (TREC method):** Judge only top documents from many systems
- **Sampling:** Only judge a subset
- **Active Learning:** System selects the most informative documents for judging
- **Reusing Judgments:** Past judgments reused for similar topics
- **Machine-Assisted Judging:** ML models predict relevance for some docs

Benefit: Greatly reduces human effort while maintaining evaluation quality.

ii) Using Statistics in Evaluation (3 Marks)

Statistics helps determine if improvements in IR systems are **significant and reliable**.

Main Techniques

- **Confidence Intervals:** Quantify reliability of metrics
- **Significance Tests:**
  - Paired t-test
  - Wilcoxon signed-rank test
    Used to confirm whether one system is truly better
- **Variance & Standard Deviation:** Measures stability across queries
- **Randomization Tests:** For comparing two ranked lists

Use Case: Checking if System A's MAP = 0.31 is *meaningfully* better than System B's MAP = 0.29.

b) Explain caching with suitable examples. [6]

Caching stores frequently accessed data in **fast memory** to reduce response time and computation.

Types of Caches in IR Systems

1) Query Result Cache: Stores results for common queries.
Example: "Cricket Score" is searched frequently—cached results return instantly.

2) Document Cache: Recently accessed documents kept in memory.
Example: Trending news articles stored for quick access.

3) Inverted List Cache: Frequently used posting lists (e.g., term "the", "India") stored in RAM to avoid disk I/O.

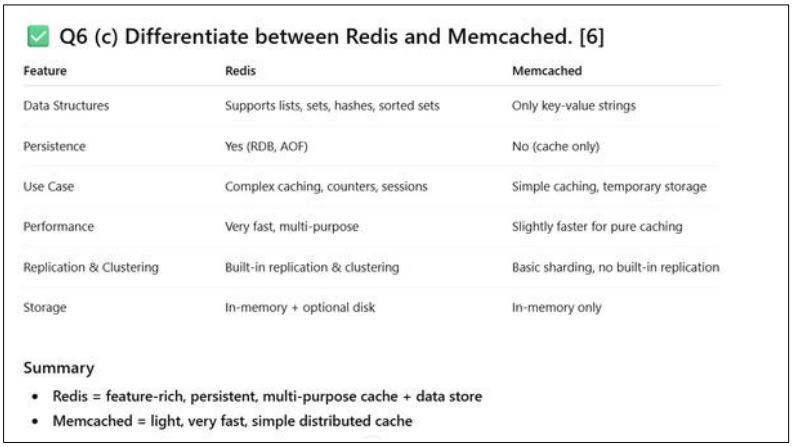
4) Cache Replacement Policies

- **LRU (Least Recently Used)**
- **LFU (Least Frequently Used)**
  Used when cache is full.

Benefits

- Reduces latency
- Reduces disk operations
- Improves system throughput
- Essential for real-time search engines like Google, Bing

c) Differentiate between Redis and Memcached. [6]



Q6 (c) Differentiate between Redis and Memcached. [6]

Feature	Redis	Memcached
Data Structures	Supports lists, sets, hashes, sorted sets	Only key-value strings
Persistence	Yes (RDB, AOF)	No (cache only)
Use Case	Complex caching, counters, sessions	Simple caching, temporary storage
Performance	Very fast, multi-purpose	Slightly faster for pure caching
Replication & Clustering	Built-in replication & clustering	Basic sharding, no built-in replication
Storage	In-memory + optional disk	In-memory only

Summary

- Redis = feature-rich, persistent, multi-purpose cache + data store
- Memcached = light, very fast, simple distributed cache

NOTE: Please verify all answers before referring.