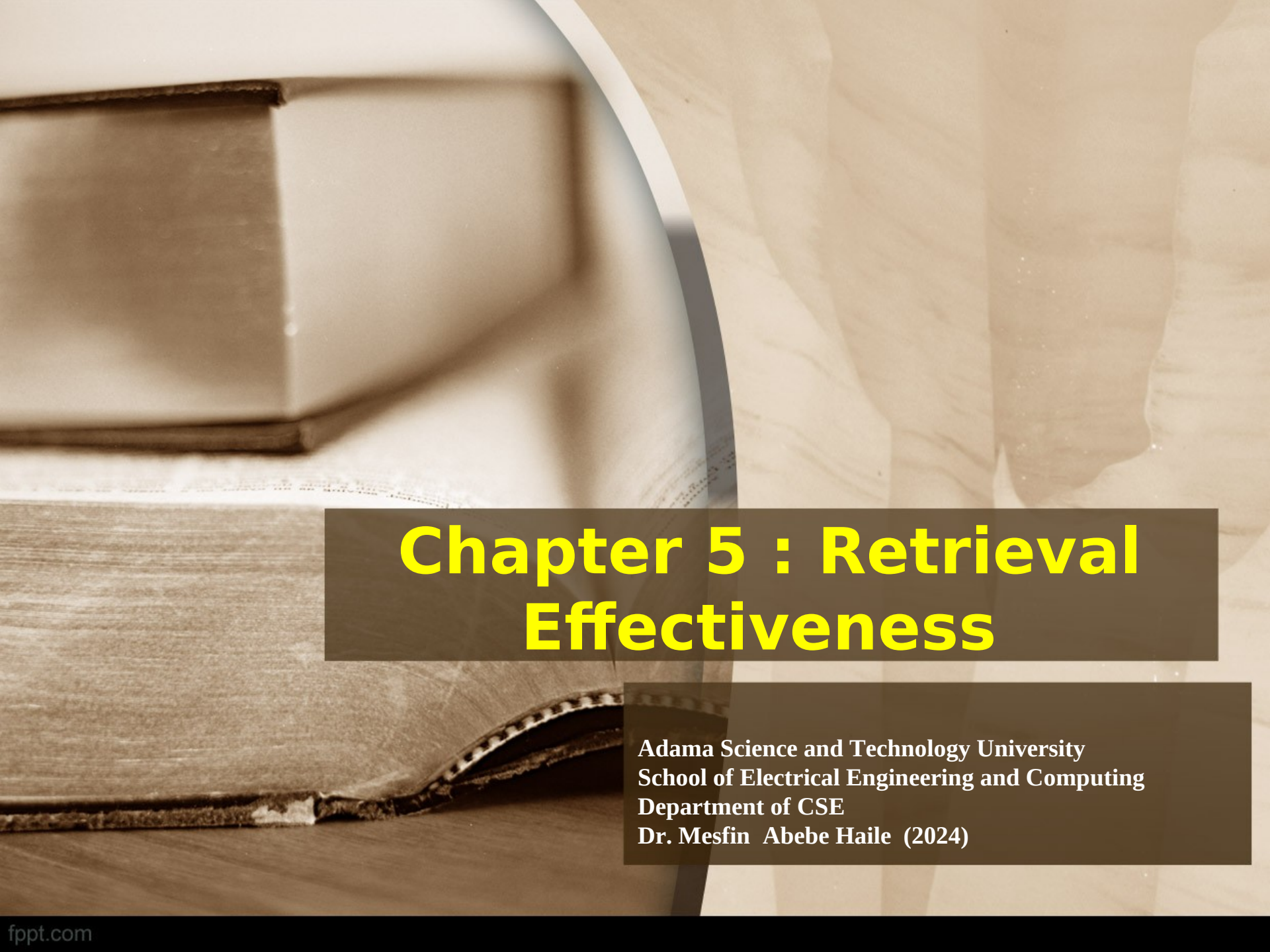




Chapter 5 : Retrieval Effectiveness

Adama Science and Technology University
School of Electrical Engineering and Computing
Department of CSE
Dr. Mesfin Abebe Haile (2024)

Retrieval Effectiveness

- ❖ Evaluation of *IR systems*,
- ❖ Relevance *judgement*,
- ❖ Performance *measures*:
 - ❖ Recall,
 - ❖ Precision,
 - ❖ Single-valued measures
 - ❖ etc.

Why System Evaluation?

- ❖ Any systems needs *validation* and *verification*.
 - ❖ Check whether the *system is right* or *not*, (Verification)
 - ❖ Check whether it is the *right system* or *not*, (Validation)
- ❖ It *provides* the ability to measure the *difference between* IR systems.
 - ❖ *How well* do our search engines work?
 - ❖ Is system A *better than* B?
 - ❖ Under *what conditions*?

Why System Evaluation?

- ❖ *Evaluation drives* what to study:
 - ❖ *Identify techniques* that work well and do not work,
 - ❖ There are *many retrieval models/algorithms*,
 - ❖ Which one is the *best*?
 - ❖ What is the *best component* for:
 - ❖ Similarity *measures* (dot-product, cosine, ...)
 - ❖ Index *term selection* (tokenization, stop-word removal, stemming...)
 - ❖ Term *weighting* (TF, TF-IDF,...)

Evaluation Criteria

- ❖ What are the *main evaluation measures* to check the *performance* of an *IR system*?
- ❖ Efficiency:
 - ❖ *Time* and *space complexity*:
 - ❖ Speed in terms of *retrieval time* and *indexing time*,
 - ❖ Speed of *query processing*,
 - ❖ The *space taken by corpus* vs. *index file*,
 - ❖ Index size: determine Index/corpus size ratio
 - ❖ Is there a need for compression?

Evaluation Criteria

❖ Effectiveness:

- ❖ How is a system capable of *retrieving relevant documents* from the *collection*?
- ❖ Is *system X better than* other systems?
- ❖ *User satisfaction*: How “*good*” are the documents that are returned as a *response to user query*?
- ❖ Relevance of results to *meet information need of users*.

Types of Evaluation Strategies

❖ System-centered evaluation:

- ❖ Given *documents*, *queries*, and *relevance judgments*,
- ❖ Try *several variations* of the *system*,
- ❖ Measure *which system returns* the “best” hit list.

❖ User-centered evaluation:

- ❖ Given *several users*, and *at least two retrieval systems*:
- ❖ Have each user try the *same task on both systems*,
- ❖ Measure *which system works* the “best” for users *information need*,
- ❖ How *to measure users satisfaction*?

The Notion of Relevance Judgment

- ❖ Relevance is the measure of a *correspondence existing between* a *document* and *query*.
 - ❖ Construct document - *query as determined by*:
 - ❖ (i) The user *who posed the retrieval problem*;
 - ❖ (ii) An *external judge*;
 - ❖ (iii) Information *specialist*.
- ❖ Is the *relevance judgment* made by *users and external person* the same ?

The Notion of Relevance Judgment

- ❖ *Relevance judgment* is usually:
 - ❖ *Subjective*: Depends upon a *specific user's judgment*.
 - ❖ *Situational*: Relates to user's *current needs*.
 - ❖ *Cognitive*: Depends on *human perception* and *behavior*.
 - ❖ *Dynamic*: Changes over time.

Measuring Retrieval Effectiveness

- Metrics often used to evaluate *effectiveness* of the system

Retrieved



Not
retrieved

Relevant	Irrelevant
A	B
C	D

“Type one error”

“Type two error”

❖ Retrieval of documents may result in:

- ❖ **False positive** (Errors of omission): some *irrelevant documents* may be retrieved by the system as *relevant*.
- ❖ **False negative** (False drop or Errors of commission): *some relevant documents* may not be retrieved by the system as *irrelevant*.
- ❖ For many applications a *good index should not permit any false drops*, but may permit a *few false positives*.

Measuring Retrieval Effectiveness

	Relevant	Not relevant
Retrieved	A	B
Not retrieved	C	D

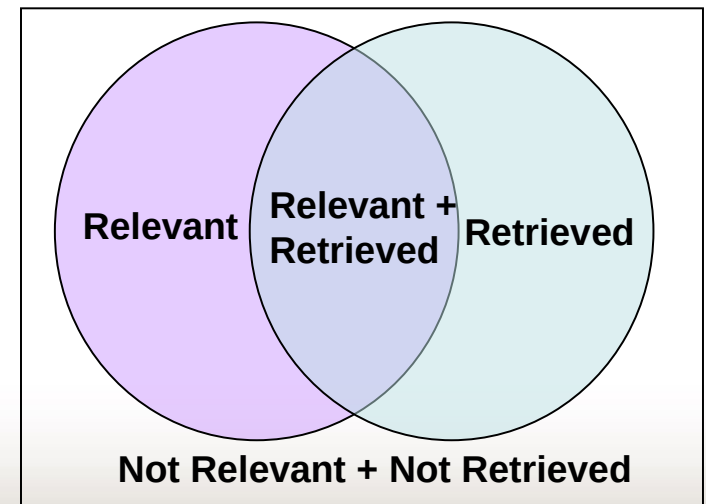
Collection size = A+B+C+D

Relevant = A+C

Retrieved = A+B

$$\text{Recall} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Relevant}\}|}$$

$$\text{Precision} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|}$$



❖ When is **precision important**? When is **recall important**?

Example

- ❖ Assume that there are a *total of 10 relevant document*.

Ranking	Relevance	Recall	Precision
1. Doc. 50	R	$0.10 = (1/10)$	$1.00 = (1/1)$
2. Doc. 34	NR	$0.10 = (1/10)$	$0.50 = (1/2)$
3. Doc. 45	R	$0.20 = (2/10)$	$0.67 = (2/3)$
4. Doc. 8	NR	$0.20 = (2/10)$	$0.50 = (2/4)$
5. Doc. 23	NR	$0.20 = (2/10)$	$0.40 = (2/5)$
6. Doc. 16	NR	$0.20 = (2/10)$	$0.33 = (2/6)$
7. Doc. 63	R	$0.30 = (3/10)$	$0.43 = (3/7)$
8. Doc 119	R	$0.40 = (4/10)$	$0.50 = (4/8)$
9. Doc 21	NR	$0.40 = (4/10)$	$0.44 = (4/9)$
10. Doc 80	R	$0.50 = (5/10)$	$0.50 = (5/10)$

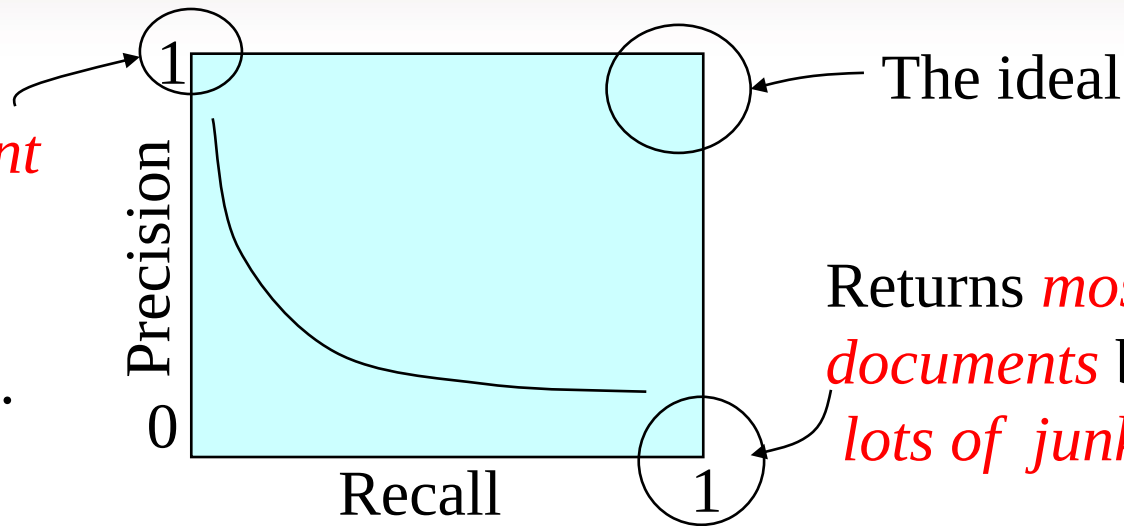
Graphing Precision and Recall

- ❖ Plot each (*recall*, *precision*) point on a graph.
- ❖ Recall is a *non-decreasing function* of the *number of documents retrieved*,
 - ❖ Precision usually *decreases* (in a *good system*)
- ❖ Precision/Recall tradeoff:
 - ❖ Can increase recall by *retrieving many documents* (*down* to a *low level* of relevance ranking),
 - ❖ But *many irrelevant documents* would be fetched, *reducing precision*.
 - ❖ Can get high recall (but low precision) by *retrieving all documents for all queries*.

Graphing Precision and Recall

- ❖ Plot each (*recall*, *precision*) point on a graph.

Returns *relevant documents* but *misses many useful* ones too.



Returns *most relevant documents* but *includes lots of junk*.

Exercise

- ❖ Let total number *of relevant documents* = 6, compute *recall* and *precision* for each cut off *point n*:

n	doc #	relevant	Recall	Precision
1	588	x	0.167	1
2	589	x	0.333	1
3	576			
4	590	x	0.5	0.75
5	986			
6	592	x	0.667	0.667
7	984			
8	988			
9	578			
10	985			
11	103			
12	591			
13	772	x	0.833	0.38
14	990			

- ❖ Missing *one relevant* document. Never reach *100% recall*.

Single-valued measures

- ❖ *Single value* measures: may want *a single value* for each query to *evaluate performance*:
 - ❖ Mean *average precision* at seen relevant documents.
 - ❖ Typically average performance over a *large set of queries*.
 - ❖ R- Precision
 - ❖ Precision at R-th relevant documents
 - ❖ F-Measure ($2*PR/P+R$)
 - ❖ E-Measure ($1-F_m$)

Problems with both precision and recall

- ❖ Number of *irrelevant documents* in the collection is not taken into account.
- ❖ **Recall** is undefined when there is *no relevant document* in the collection. **Precision** is undefined when *no document is retrieved*.

Other measures

- ❖ **Noise** = retrieved irrelevant docs / retrieved docs.
- ❖ **Silence/Miss** = non-retrieved relevant docs / relevant docs.
 - ❖ **Noise** = 1 – Precision; **Silence** = 1 – Recall

$$Miss = \frac{|\{Relevant\} \cap \{NotRetrieved\}|}{|\{Relevant\}|}$$

$$Fallout = \frac{|\{Retrieved\} \cap \{NotRelevant\}|}{|\{NotRelevant\}|}$$

Programming Assignment

- ❖ Select the *language that interest* you and *design an IR system* for document collection written in that language.
- ❖ Form a group of not more than three members
- 1. *Construct indexing* structure:
 - ❖ Given text document collection generate index terms and organize them using inverted file indexing, include TF, DF & CF for each index term and position/location information of terms in each document.
- 1. *Develop Vector space retrieval* model:
 - ❖ Using Vector space model generate any two queries (with two or more words each) and retrieve relevant documents in ranking order.

Question & Answer



Thank You !!!