Informatics 225
Computer Science 221

Information Retrieval

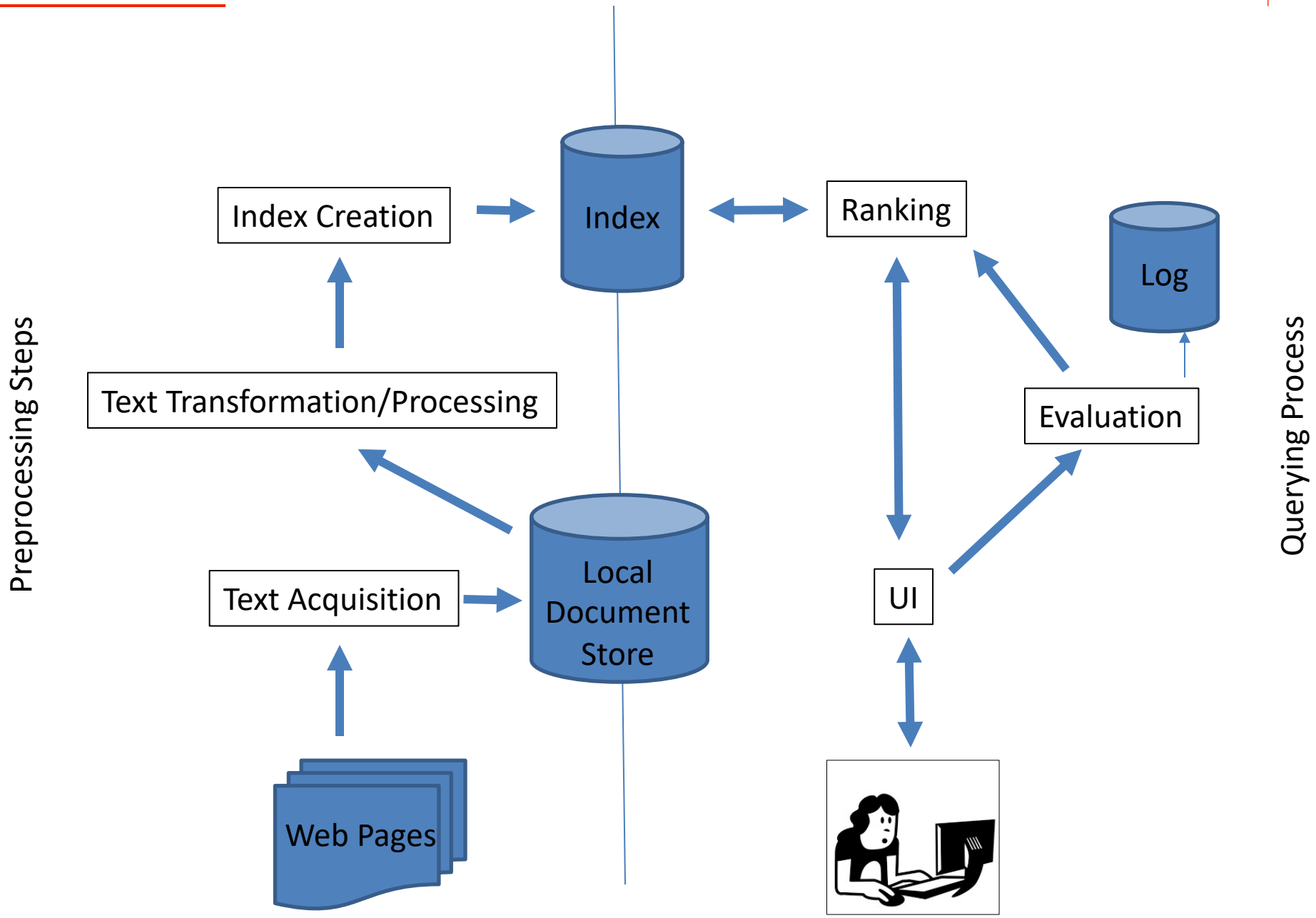
Lecture 25

These course materials borrow, with permission, from those of Prof. Cristina Videira Lopes, Addison Wesley 2008, Chris Manning, Pandu Nayak, Hinrich Schütze, Heike Adel, Sascha Rothe, Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. Powerpoint theme by Prof. André van der Hoek.

Search Engine Evaluation

Information Retrieval

Architecture



Evaluation

- For a given query, a corpus and a specific definition of relevance:

Evaluation

- For a given query, a corpus and a specific definition of relevance:

  Effectiveness and *efficiency*

  - Effectiveness : measures the ability to find the right information
    - How well the rank corresponds to the rank that the user expects

Evaluation

- For a given query, a corpus and a specific definition of relevance:

  Effectiveness and *efficiency*

  - Effectiveness : measures the ability to find the right information
    - How well the rank corresponds to the rank that the user expects

  - Efficiency : measures how quickly this is done, how many resources are needed for this to be done
    - Time and space requirements of the methods that produced the ranking

Evaluation

- Evaluation is key to building *effective* and *efficient* search engines
  - measurement usually carried out in controlled laboratory experiments
  - *online* testing can also be done

Evaluation

- Evaluation is key to building *effective* and *efficient* search engines
  - measurement usually carried out in controlled laboratory experiments
  - *online* testing can also be done

- Effectiveness, efficiency and *cost* are related
  - e.g., if we want a particular level of effectiveness and efficiency, this will determine the cost of the system configuration
  - efficiency and cost targets may impact effectiveness

Evaluation Corpus

- Allow results from different methods to be compared

Evaluation Corpus

- Allow results from different methods to be compared

- *Test collections* consisting of documents, queries, and relevance judgments, e.g.,

  - CACM: Titles and abstracts from the Communications of the ACM from 1958-1979. Queries and relevance judgments generated by computer scientists.

Evaluation Corpus

- Allow results from different methods to be compared

- *Test collections* consisting of documents, queries, and relevance judgments, e.g.,

  - CACM: Titles and abstracts from the Communications of the ACM from 1958-1979. Queries and relevance judgments generated by computer scientists.

  - AP: Associated Press newswire documents from 1988-1990 (from TREC disks 1-3). Queries are the title fields from TREC topics 51-150. Topics and relevance judgments generated by government information analysts.

  - GOV2: Web pages crawled from websites in the .gov domain during early 2004. Queries are the title fields from TREC topics 701-850. Topics and relevance judgments generated by government analysts.

Test Collections

Collection	Number of documents	Size	Average number of words/doc.
CACM	3,204	2.2 Mb	64
AP	242,918	0.7 Gb	474
GOV2	25,205,179	426 Gb	1073

Collection	Number of queries	Average number of words/query	Average number of relevant docs/query
CACM	64	13.0	16
AP	100	4.3	220
GOV2	150	3.1	180

TREC Topic Example for a query

<top>
<num> Number: 794

<title> pet therapy

The title is used as query

<desc> Description:
How are pets or animals used in therapy for humans and what are the benefits?

<narr> Narrative:
Relevant documents must include details of how pet- or animal-assisted therapy is or has been used. Relevant details include information about pet therapy programs, descriptions of the circumstances in which pet therapy is used, the benefits of this type of therapy, the degree of success of this therapy, and any laws or regulations governing it.

</top>

Relevance Judgments

- Obtaining relevance judgments is an expensive, time-consuming process
  - who does it?
  - what are the instructions?
  - what is the level of agreement?

Relevance Judgments

- Obtaining relevance judgments is an expensive, time-consuming process
  - who does it?
  - what are the instructions?
  - what is the level of agreement?

- TREC judgments
  - depend on task being evaluated
  - generally binary
  - agreement good because of “narrative”

Pooling

- Exhaustive judgments for all documents in a collection is not practical

Pooling

- Exhaustive judgments for all documents in a collection is not practical

- Pooling technique is used in GOV2
  - top *k results (for TREC, k varied between 50 and* 200) from the rankings obtained by different search engines (or retrieval algorithms) are merged into a pool
  - duplicates are removed
  - documents are presented in some random order to the relevance judges

Pooling

- Exhaustive judgments for all documents in a collection is not practical

- Pooling technique is used in GOV2
  - top *k results (for TREC, k varied between 50 and* 200) from the rankings obtained by different search engines (or retrieval algorithms) are merged into a pool
  - duplicates are removed
  - documents are presented in some random order to the relevance judges

- Produces a large number of relevance judgments for each query, although still incomplete
  - Studies have shown that comparisons are accurate

Query Logs

- Used for both tuning and evaluating search engines
  - also for various techniques such as query suggestion

Query Logs

- Used for both tuning and evaluating search engines
  - also for various techniques such as query suggestion
- Can generate privacy concerns
  - Anonymize the logged data?

Query Logs

- Used for both tuning and evaluating search engines
  - also for various techniques such as query suggestion
- Can generate privacy concerns
  - Anonymize the logged data?
- Typical contents
  - User identifier or user session identifier
  - Query terms - stored exactly as user entered
  - List of URLs of results, their ranks on the result list, and whether they were clicked on
  - Timestamp(s) - records the time of user events such as query submission, clicks
  - In which page the user Clicked!!!

Query Logs

- Clicks are not relevance judgments
  - although they are correlated
  - biased by a number of factors such as rank on result list: confirmation bias!

Query Logs

- Clicks are not relevance judgments
  - although they are correlated
  - biased by a number of factors such as rank on result list: confirmation bias!

- Can use clickthough data to predict *preferences* between pairs of documents
  - Correlated with relevance judgments
  - Appropriate for tasks with multiple levels of relevance, focused on user relevance
  - Various “policies” used to generate preferences

Example Click Policy

- *Skip Above and Skip Next* (Agichtein et al., 2006)
  - Result document list and click data

$$d_1$$
$$d_2$$
$$d_3 \ (\text{clicked})$$
$$d_4$$

  - Generated preferences

$$d_3 > d_2$$
$$d_3 > d_1$$
$$d_3 > d_4$$

Query Logs

- Click data of a single user can be noisy!
  - But click data can also be aggregated to remove noise

Query Logs

- Click data of a single user can be noisy!
  - But click data can also be aggregated to remove noise

- *Click distribution* information
  - can be used to identify clicks that have a higher frequency than would be expected
  - high correlation with relevance
  - e.g., using *click deviation* to filter clicks for preference-generation policies

Filtering Clicks

- *Click deviation CD(d, p)* for a result *d* in position *p*:

$$CD(d,p) = O(d,p) - E(p)$$

O(d,p): observed click frequency for a document in a rank position p *over all instances of a given query*

E(p): expected click frequency at rank p *averaged across all queries*

- Use this value to filter clicks and provide more reliable information (optimizing your ranking).

Effectiveness Measures

A is set of relevant documents,
B is set of retrieved documents

Recall *: how well the search engine is doing at finding all the relevant documents for a query.*
Precision *: how well it is doing at rejecting non-relevant documents.*

Effectiveness Measures

A is set of relevant documents,
B is set of retrieved documents

	Relevant	Non-Relevant
Retrieved	$A \cap B$	$\overline{A} \cap B$
Not Retrieved	$A \cap \overline{B}$	$\overline{A} \cap \overline{B}$

$$Recall = \frac{|A \cap B|}{|A|}$$

$$Precision = \frac{|A \cap B|}{|B|}$$

Recall : how well the search engine is doing at finding all the relevant documents for a query.
Precision : how well it is doing at rejecting non-relevant documents.

Classification Errors

- *False Positive* (Type I error)
  - a non-relevant document is retrieved

$$Fallout = \frac{|\overline{A} \cap B|}{|\overline{A}|}$$

Classification Errors

- *False Positive* (Type I error)
  - a non-relevant document is retrieved

$$Fallout = \frac{|\overline{A} \cap B|}{|\overline{A}|}$$

- *False Negative* (Type II error)
  - Relevant documents that are not retrieved
  - 1- *Recall*