ADVANCED TOPICS IN INFORMATION RETRIEVAL AND WEB SEARCH

Lecture 4:

Information Retrieval Evaluation

Dr. S. Mehdi Vahidipour
Thanks Dr. Momtazi

Vahidipour@kashanu.ac.ir

Based on the text book slides.

Outline

- Introduction
- Unranked vs ranked retrieval
  - Evaluation of unranked retrieval
  - Evaluation of ranked retrieval
- **Significant tests**
- Evaluation data and benchmarks
- Evaluation at large scale data
- Results representation

Significance Tests

- Given the results from a number of queries, how can we conclude that ranking algorithm A is better than algorithm B?
  - Using an effectiveness metric: significant test
- 2 sets of data with normal distributions
- A difference is considered significant if the probability of getting that difference by random chance is very small

Hypothesis Tests

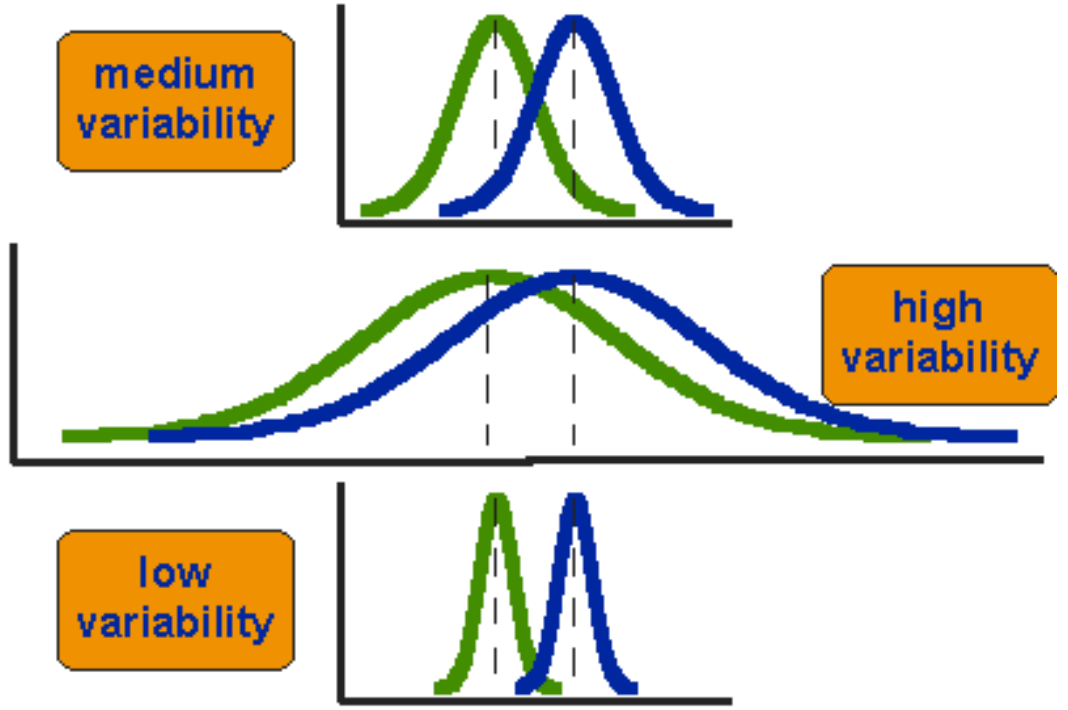
- Hypothesis:
  - A statement which can be proven false
- Null hypothesis (HO):
  - “There is no difference”
- Alternative hypothesis (HA):
  - “There is a difference…”
- Try to “reject the null hypothesis”
  - If the null hypothesis is false, it is likely that our alternative hypothesis is true
  - “False” – there is only a small probability that the results we observed could have occurred by chance

Significance Tests

- A significance test enables us to reject the null hypothesis (“no difference”) in favor of the alternative hypothesis (“B is better than A”)
  - A is baseline, B is “new and improved” version
  - The power of a test is the probability that the test will reject the null hypothesis correctly
  - Increasing the number of queries in the experiment also increases power of test

t-Test

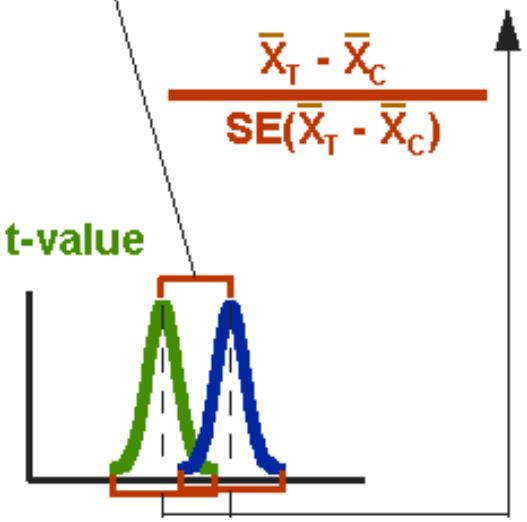
- Comparing the means of two groups is not enough
- We need to consider the whole distribution



t-test

- The formula for t-test is a ratio (signal-to-noise ratio)
  - Top part: difference between the two means or averages
  - Bottom part: a measure of the variability or dispersion of the scores

$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}} = \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)} = \text{t-value}$$



$$t = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}}$$

t-test

- By computing the t-value, we have to look it up in the table of significance to test whether the ratio is large enough to say that the difference between the groups is not likely to have been a chance finding
- P value:
  - The probability of making an error by chance
  - Historically we use $p < 0.05$
- Two tail vs one tail

Critical Values for One-sided and Two-sided Tests Using Student's t Distribution

1 tail:	0.25	0.1	0.05	0.025	0.01	0.005	0.001
2 tail:	0.5	0.2	0.1	0.05	0.02	0.01	0.002
1	1.000	3.078	6.314	12.706	31.821	63.657	318.309
2	0.816	1.886	2.920	4.303	6.965	9.925	22.327
3	0.765	1.638	2.353	3.182	4.541	5.841	10.215
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930

t-test

- Example:
  - $\overline{x_1} = 13.0$
  - $\overline{x_2} = 11.89$
  - $var_1 = 15.11$
  - $var_2 = 16.61$
  - $t = \frac{13.0-11.89}{\sqrt{\frac{15.11}{10}+\frac{16.61}{9}}} = 0.61$
  - $Degree\ of\ Freeom\ df = n_1 + n_2 - 2 = 17$

- Minimum t-value in lookup table

  for df=17 and p<0.05: 2.11

  => the difference is not statistically significant

Group 1 ($N_1 = 10$)	Group 2 ($N_2 = 9$)
18	13
15	14
13	12
17	6
14	11
8	13
10	17
11	16
7	5
17	

Outline

- Introduction
- Unranked vs ranked retrieval
  - Evaluation of unranked retrieval
  - Evaluation of ranked retrieval
- Significant tests
- **Evaluation data and benchmarks**
- Evaluation at large scale data
- Results representation

Evaluation Data

- Goals
  - Provide fixed experimental setting and data
  - Ensure fair and repeatable experiments
- Text corpus is without queries and relevance judgment
  - Linguistics, machine translation, speech recognition
- Benchmarks includes test collection of
  - Documents
  - Queries
  - Relevance judgments

Evaluation Data

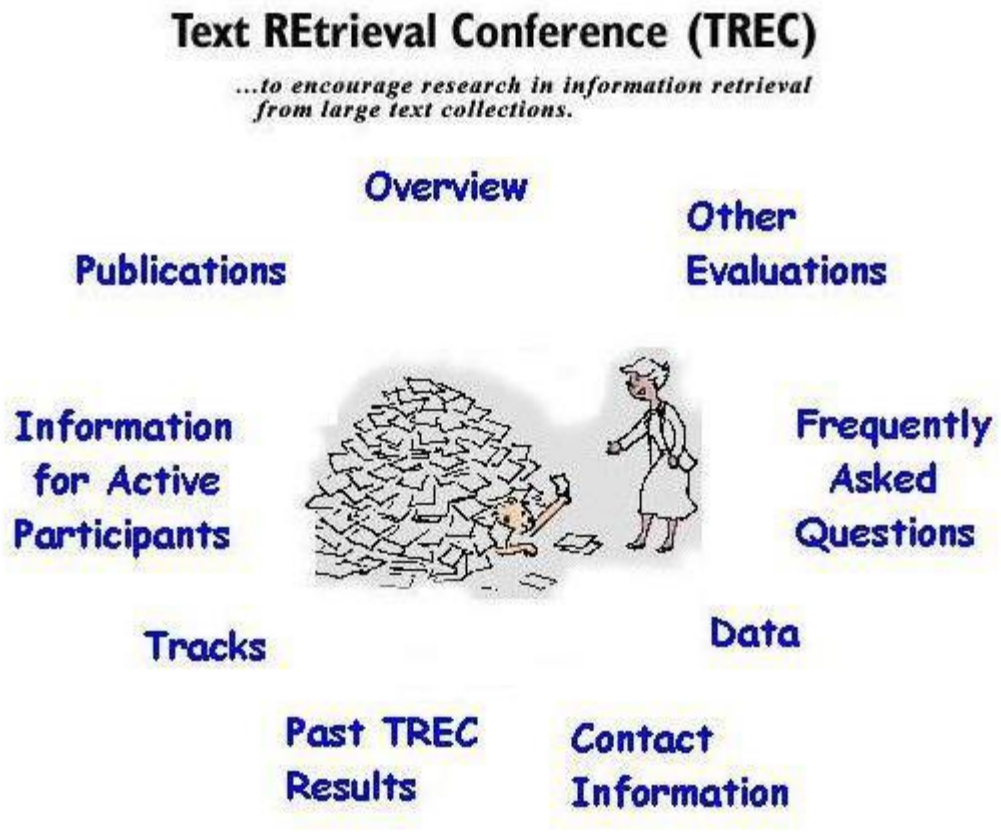
- A collection of documents
  - Documents should be representative of the documents we expect to see in reality.
- A collection of information needs (often incorrectly called queries)
  - Information needs should be representative of the information needs we expect to see in reality.
- Human relevance assessments
  - We need to hire/pay “judges” or assessors to do this.
  - Expensive, time-consuming
  - Judges should be representative of the users we expect to see in reality.

First Standard Relevance Benchmark: Cranfield

- Pioneering: first testbed allowing precise quantitative tests
- Measures of information retrieval effectiveness
- Late 1950s, UK
- 1398 abstracts of aerodynamics journal articles, a set of 225 queries, exhaustive relevance judgments of all query-document-pairs
- Too small, too untypical for serious IR evaluation today

Second-generation Relevance Benchmark: TREC

- TREC = Text Retrieval Conference (TREC)
- Organized by the U.S. National Institute of Standards and Technology (NIST)
- TREC is actually a set of several different relevance benchmarks.
- Best known: TREC Ad Hoc, used for first 8 TREC evaluations between 1992 and 1999



Second-generation Relevance Benchmark: TREC

- TREC Ad Hoc task (1992 and 1999)
  - 1.89 million documents, mainly newswire articles, 450 information needs
  - No exhaustive relevance judgments – too expensive
  - Result assessment: NIST assessors' relevance judgments are available only for the documents that were among the top k returned for some system which was entered in the TREC evaluation for which the information need was developed.

More Recent Benchmark: ClueWeb09

- 1 billion web pages
- 25 terabytes (compressed: 5 terabyte)
- Collected January/February 2009
- 10 languages
- Unique URLs: 4,780,950,903 (325 GB uncompressed, 105 GB compressed)
- Total Outlinks: 7,944,351,835 (71 GB uncompressed, 24 GB compressed)

Evaluation Corpora

- Corpora change (in particular grow) over time
- 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 Web sites 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

Validity of Relevance Assessments

- Relevance assessments are only usable if they are consistent
- If they are not consistent, then there is no “truth” and experiments are not repeatable

- Measuring the consistency or agreement among judges

→ Kappa measure

Kappa Measure

- Kappa is measure of how much judges agree or disagree.
- Designed for categorical judgments and corrected for chance agreement

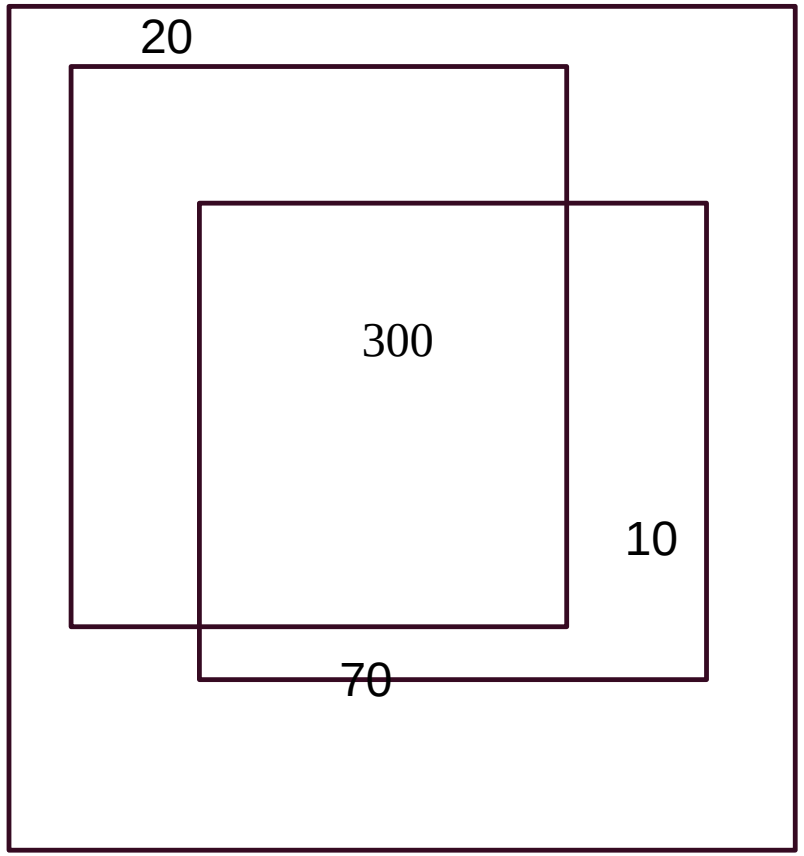
- P(A) = proportion of time judges agree
- P(E) = what agreement would we get by chance

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

- Values of κ in the interval [2/3, 1.0] are seen as acceptable.
  - $\kappa > 0.8$: good agreement
  - $0.67 < \kappa < 0.8$: "tentative conclusions"
  - With smaller values: need to redesign relevance assessment methodology used etc

Kappa Measure

- Example:
  - Total data: 400
  - Annotator 1: 320 relevant & 80 nonrelevant
  - Annotator 2: 310 relevant & 90 nonrelevant
  - Overlap: 300 relevant & 70 nonrelevant



Kappa Measure

- Observed proportion of the times the judges agreed
  - $P(A) = \frac{300+70}{400} = \frac{370}{400} = 0.925$
- Probability that the two judges agreed by chance
  - $P(E) = P(nonrelevant)^2 + P(relevant)^2$
  - $P(nonrelevant) = \frac{80+90}{400+400} = \frac{170}{800} = 0.2125$
  - $P(relevant) = \frac{320+310}{400+400} = \frac{680}{800} = 0.7878$
  - $P(E) = P(nonrelevant)^2 + P(relevant)^2 = 0.2125^2 + 0.7878^2$
- Kappa statistic:

$$\kappa = \frac{P(A\_-P(E)}{(1-P(E)} = \frac{.925-.665}{1-.665} = .776\text{(in acceptable range)}$$

Outline

- Introduction
- Unranked vs ranked retrieval
  - Evaluation of unranked retrieval
  - Evaluation of ranked retrieval
- Significant tests
- Evaluation data and benchmarks
- **Evaluation at large scale data**
- Results representation

Evaluation at Large Search Engines

- Recall is difficult to measure on the web
- Search engines often use precision at top k, e.g., k = 10 . . .
- . . . or use measures that reward you more for getting rank 1 right than for getting rank 10 right.
- Search engines also use non-relevance-based measures
  - Example: clickthrough on first result

Clickthrough Data

- It can be obtained by observing how frequently the users click on a given document, when it is shown in the answer set for a given query
- Clicks are not relevance judgments
- … although they are highly correlated
- Not very reliable if you look at a single clickthrough
  - You may realize after clicking that the summary was misleading and the document is nonrelevant . . .
  - . . . but pretty reliable in the aggregate
- This is particularly attractive because the data can be collected at a low cost without overhead for the user

Clickthrough Data

- Clickthrough data is difficult to interpret
- Biased by a number of factors:
  - Rank on result list
  - Snippet
  - General popularity
- Other indicators
  - Dwell time: time spent on a clicked result
  - Search exit action: result page, print page, timeout, enter other URL, …

Outline

- Introduction
- Unranked vs ranked retrieval
  - Evaluation of unranked retrieval
  - Evaluation of ranked retrieval
- Significant tests
- Evaluation data and benchmarks
- Evaluation at large scale data
- **Results representation**

Results Representation

- How do we present results to the user?
- Most often: as a list; e.g., “10 blue links”

- How should each document in the list be described?
  - This description is crucial
  - The user often can identify good hits (= relevant hits) based on the description
  - No need to actually view any document

- Most commonly: doc title, url, some metadata . . .
- . . . and a summary

Documents' Summaries

- Two basic kinds:
  - Static
    - A static summary of a document is always the same, regardless of the query that was issued by the user.
  - Dynamic
    - Dynamic summaries are query-dependent. They attempt to explain why the document was retrieved for the query at hand.

Static Summaries

- In typical systems, the static summary is a subset of the document.
- Simplest heuristic: the first 50 or so words of the document
- More sophisticated: extract from each document a set of "key" sentences
  - Simple NLP heuristics to score each sentence
  - Summary is made up of top-scoring sentences
  - Machine learning approach: see IIR 13
- Most sophisticated: complex NLP to synthesize/generate a summary
  - For most IR applications: not quite ready for prime time yet

Dynamic Summaries

- Present one or more “windows” or snippets within the document that contain several of the query terms.
- Prefer snippets in which query terms occurred as a phrase
- Prefer snippets in which query terms occurred jointly in a small window
- The summary that is computed this way gives the entire content of the window – all terms, not just the query terms.
- Criteria:
  - Occurrence of keywords, density of keywords, coherence of snippet, number of different snippets in summary, good cutting points etc

Dynamic Summaries

- Example

Dynamic Summaries

- Tradeoff between short and long snippets
  - Snippets must be short, since real estate on the search result page is limited
  - Snippets must be long enough to be meaningful

- Snippets should communicate whether and how the document answers the query
- Ideally: linguistically well-formed snippets
- Ideally: the snippet should answer the query, so we don't have to look at the document.
- Dynamic summaries are a big part of user happiness because …
  - . . . we can quickly scan them to find the relevant document we then click on
  - . . . in many cases, we don't have to click at all and save time

Questions?