Introduction to Information Retrieval

3. seminar

Measuring relevance effectiveness

University of Pannonia

Tamás Kiezer, Miklós Erdélyi

About relevance effectiveness

- Motivation
  - IR system development
  - Evaluating/comparing IR systems
- Definition of relevance effectiveness:
  - “the ability of a retrieval method or system to return relevant answers”
  - eg., how well or bad an IR system performs
- Relevance is subjective!

IR system evaluation

- Choose measurement method(s)
- Choose data to use for evaluation (eg. a standard test collection)
- Measure relevance effectiveness according to the chosen method(s)

Standard test collections

- Well-controlled
- Contain documents, queries, and relevance assessments for (most) query-document pairs
- Example:
  - CRAN (1950s)
  - TREC (Text REtrieval Conference, from 1992)
  - ADI
- Question: What could be problematic with measuring a Web search engine?

Sets-based measures (1)

- Precision (P): #(relevant items retrieved) / #(retrieved items)
- Recall (R): #(relevant items retrieved) / #(relevant items)
- Fallout: [#(retrieved items) - #(relevant items retrieved)] / [#total items) - #(relevant items)]
- Combination:
  - F-measure: trades off precision and recall
  - F = 2PR/(P+R)

Sets-based measures (2)

- Easily visualized by contingency table:

	relevant	nonrelevant
retrieved	true positives (tp)	false positives (fp)
not retrieved	false negatives (fn)	true negatives (tn)

- Then:
  - P = tp/(tp + fp)
  - R = tp/(tp + fn)
- Other uses…

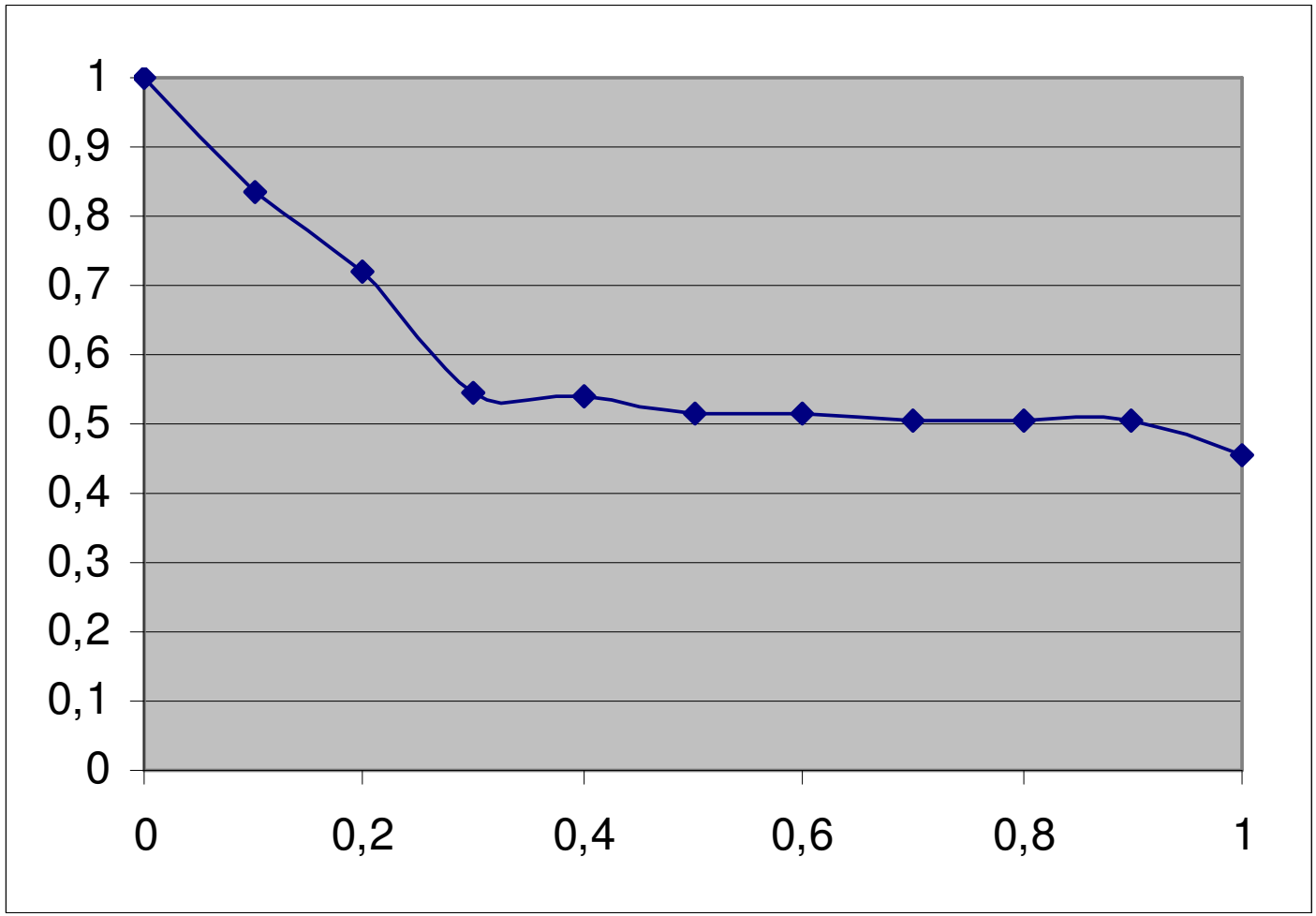
Exercise: plotting a PR graph, calculating MAP

- In response to queries q_1, q_2, q_3 an IR system returned the following sets of documents (relevant ones are starred) out of 125 documents:

$$\mathrm{MAP}(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\mathrm{Precision}(R_{jk})$$

	$\mathbf{q_1}$	$\mathbf{q_2}$	$\mathbf{q_3}$
1.	d_{15}	d_{17} *	d_5 *
2.	d_{23} *	d_3	d_{43}
3.	d_5	d_9 *	d_{54} *
4.	d_{76} *	d_{16}	d_{36}
5.	d_{125}	d_{25} *	d_{41} *
6.	d_{43}	d_{31}	d_{51}
7.	d_{84}	d_{48} *	d_{63} *
8.	d_6 *		d_{26}
9.	d_{24}		
10.	d_{89}		
	Δ=15	Δ=10	Δ=6

Solution: precision-recall graph



Evaluation of ranked retrieval sets

- Results are *ranked*, ie., ordered
- Methods:
  - [MAP (mean average precision)]
  - RP
  - M-L-S (user-based)

Review: RP method

1. Select meta-search engine to be measured.
2. Define queries q_i, $i = 1,\ldots,n$.
3. Define the value of m; typically $m = 5$ or $m = 10$.
4. Perform searches for every q_i using the meta-search engine
   as well as the search engines used by the meta-search engine, $i = 1,\ldots,n$.
5. Compute relative precision **for** q_i as follows:

$$RP_{q_i,m} = \frac{T_i}{V_i}, \quad i = 1,\ldots,n$$

6. Compute average: $\sum_{i=1}^{n} RP_{q_i,m}$ **/n**

Exercise: RP method

- Task: compute the average RP of MetaCrawler.com for the given queries.
- Search engines: Google, Yahoo!, MSN
- Settings:
  - m=5
  - q_1= “strange museums”
  - q_2= “free wallpapers”

Review: M-L-S method

1. Select search engine to be measured.
2. Define relevance categories.
3. Define groups.
4. Define weights.
5. Give queries q_i $(i = 1,...,s)$.
6. Compute $P5_i$ and/or $P10_i$ for q_i $(i=1,...,s)$.
7. The first 5/10-precision of the search engine is:

$$Pk = \frac{1}{s}\sum_{i=1}^{s} Pk_i \text{ , where } k = 5 \text{ or } k = 10.$$

Exercise: M-L-S method

- Task: compute first 5 precision for the given hit lists of Google and Yahoo! according to *your* relevance judgement, and compare the two search engines.
- Categories: relevant/not relevant
- Groups:
  - First two hits, next three hits
- Weights:
  - First group: 10, second group: 5
- Queries:
  - q_1= “gallup”
  - q_2= “kosár”

$$P5 = \frac{no\_relevant\_hits_{1.-2.hit} \times 10 + no\_relevant\_hits_{3.-5.hit} \times 5}{35 - ((5 - no\_hits_{1.-5.hit}) \times 5)}$$

Questions?