Today

1. General concepts in IR evaluation
2. The TREC competitions
3. IR evaluation metrics
Evaluation: difficulties

- IR system
  - in: a query
  - out: relevant documents

- Evaluation of IR systems

- Goal: predict future from past experience

- Reasons why IR evaluation is hard:
  - Large variation in human information needs and queries
  - The precise contributions of each component are hard to entangle:
    - Collection coverage
    - Document indexing
    - Query formulation
    - Matching algorithm

Evaluation: “the laboratory model”

- Test only “system parameters”
  - Index language devices for description and search
  - Methods of term choice for documents
  - Matching algorithm
  - Type of user interface

- Ignore environment variables
  - Properties of documents → use many documents
  - Properties of users → use many queries
What counts as acceptable test data?

- In 60s and 70s, very small test collections, arbitrarily different, one per project
  - in 60s: 35 queries on 82 documents
  - in 1990: still only 35 queries on 2000 documents
- not always kept test and training apart as so many environment factors were tested
- TREC-3: 742,000 documents; TREC Web-track: small snapshot of the web
- Large test collections are needed
  - to capture user variation
  - to support claims of statistical significance in results
  - to demonstrate that systems scale up → commercial credibility
- Practical difficulties in obtaining data; non-balanced nature of the collection

Today’s test collections

A test collection consists of:

- Document set:
  - Large, in order to reflect diversity of subject matter, literary style, noise such as spelling errors
- Queries/Topics
  - short description of information need
  - TREC “topics”: longer description detailing relevance criteria
  - “frozen” → reusable
- Relevance judgements
  - binary
  - done by same person who created the query
- Text REtrieval Conference
- Run by NIST (US National Institute of Standards and Technology)
- Marks a new phase in retrieval evaluation
  - common task and data set
  - many participants
  - continuity
- Large test collection: text, queries, relevance judgements
  - Queries devised and judged by information specialist (same person)
  - Relevance judgements done only for up to 1000 documents/query
- 2003 was 12th year
- 87 commercial and research groups participated in 2002

Sample TREC query

<num> Number: 508
<title> hair loss is a symptom of what diseases
<desc> Description:
Find diseases for which hair loss is a symptom.
<narr> Narrative:
A document is relevant if it positively connects the loss of head hair in humans with a specific disease. In this context, "thinning hair" and "hair loss" are synonymous. Loss of body and/or facial hair is irrelevant, as is hair loss caused by drug therapy.
Humans decide which document–query pairs are relevant.

Evaluation metrics

<table>
<thead>
<tr>
<th></th>
<th>Relevant</th>
<th>Non-relevant</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved</td>
<td>A</td>
<td>B</td>
<td>A+B</td>
</tr>
<tr>
<td>Not retrieved</td>
<td>C</td>
<td>D</td>
<td>C+D</td>
</tr>
<tr>
<td>Total</td>
<td>A+C</td>
<td>B+D</td>
<td>A+B+C+D</td>
</tr>
</tbody>
</table>

Recall: proportion of retrieved items amongst the relevant items \(\frac{A}{A+C}\)

Precision: proportion of relevant items amongst retrieved items \(\frac{A}{A+B}\)

Accuracy: proportion of correctly classified items as relevant/irrelevant \(\frac{A+D}{A+B+C+D}\)

Recall: [0..1]; Precision: [0..1]; Accuracy: [0..1]

Accuracy is not a good measure for IR, as it conflates performance on relevant items (A) with performance on irrelevant (uninteresting) items (D)
Recall and Precision

- All documents: 
  \[ A+B+C+D = 130 \]

- Relevant documents for a given query: 
  \[ A+C = 28 \]

Recall and Precision: System 1

- System 1 retrieves 25 items: 
  \[ (A+B)_1 = 25 \]

- Relevant and retrieved items: 
  \[ A_1 = 16 \]

\[
R_1 = \frac{A_1}{A+C} = \frac{16}{28} = .57
\]
\[
P_1 = \frac{A_1}{(A+B)_1} = \frac{16}{25} = .64
\]
\[
A_1 = \frac{A_1+D_1}{A+B+C+D} = \frac{16+93}{130} = .84
\]
Recall and Precision: System 2

- System B retrieves set \((A+B)_2 = 15\) items
- \(A_2 = 12\)

\[R_2 = \frac{12}{28} = .43\]
\[P_2 = \frac{12}{15} = .8\]
\[A_2 = \frac{12+99}{130} = .85\]

Recall-precision curve

- Plotting precision and recall (versus no. of documents retrieved) shows inverse relationship between precision and recall
- Precision/recall cross-over can be used as combined evaluation measure

- Plotting precision versus recall gives recall-precision curve
- Area under normalised recall-precision curve can be used as evaluation measure
Inverse relationship between precision and recall forces general systems to go for compromise between them.

But some tasks particularly need good precision whereas others need good recall:

<table>
<thead>
<tr>
<th>Precision-critical task</th>
<th>Recall-critical task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Little time available</td>
<td>Time matters less</td>
</tr>
<tr>
<td>A small set of relevant documents answers the information need</td>
<td>One cannot afford to miss a single document</td>
</tr>
<tr>
<td>Potentially many documents might fill the information need (redundantly)</td>
<td>Need to see each relevant document</td>
</tr>
<tr>
<td>Example: web search for factual information</td>
<td>Example: patent search</td>
</tr>
</tbody>
</table>

The problem of determining recall

- Recall problem: for a collection of non-trivial size, it becomes impossible to inspect each document.
- It would take 6500 hours to judge 800,000 documents for one query (30 sec/document).
- Pooling addresses this problem.
Pooling (Sparck Jones and van Rijsbergen, 1975)

- Pool is constructed by putting together top $N$ retrieval results from a set of $n$ systems (TREC: $N = 100$)
- Humans judge every document in this pool
- Documents outside the pool are automatically considered to be irrelevant
- There is overlap in returned documents: pool is smaller than theoretical maximum of $N \cdot n$ systems (around $\frac{1}{3}$ the maximum size)
- Pooling works best if the approaches used are very different
- Large increase in pool quality by manual runs which are recall-oriented, in order to supplement pools

F-measure

- Weighted harmonic mean of P and R (Rijsbergen 1979)

\[ F_\alpha = \frac{PR}{(1 - \alpha)P + \alpha R} \]

  - High $\alpha$: Precision is more important
  - Low $\alpha$: Recall is more important

- Most commonly used with $\alpha = 0.5 \rightarrow$

\[ F_{0.5} = \frac{2PR}{P + R} \]

- Maximum value of $F_{0.5}$-measure (or F-measure for short) is a good indication of best P/R compromise
- F-measure is an approximation of cross-over point of precision and recall
• With ranked list of return documents there are many P/R data points
• Sensible P/R data points are those after each new relevant document has been seen (black points)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Relev.</th>
<th>R</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>0.20</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>0.40</td>
<td>0.67</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td>0.60</td>
<td>0.50</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>X</td>
<td>0.80</td>
<td>0.40</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>X</td>
<td>1.00</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Summary IR measures

• Precision at a certain rank: P(100)
• Precision at a certain recall value: P(R=.2)
• Precision at last relevant document: P(last_relev)
• Recall at a fixed rank: R(100)
• Recall at a certain precision value: R(P=.1)
• Want to average over queries
• Problem: queries have differing number of relevant documents
• Cannot use one single cut-off level for all queries
  – This would not allow systems to achieve the theoretically possible maximal values in all conditions
  – Example: if a query has 10 relevant documents
    * If cutoff > 10, $P < 1$ for all systems
    * If cutoff < 10, $R < 1$ for all systems
• Therefore, more complicated joint measures are required

### 11 point average precision

$$P_{11\text{-}pt} = \frac{1}{11} \sum_{j=0}^{10} \frac{1}{N} \sum_{i=1}^{N} \tilde{P}_i(r_j)$$

with $\tilde{P}_i(r_j)$ the precision at the $j$th recall point in the $i$th query (out of $N$ queries)

• Define 11 standard recall points $r_j = \frac{j}{10}$: $r_0 = 0$, $r_1 = 0.1$ ... $r_{10} = 1$
• We need $\tilde{P}_i(r_j)$; i.e. the precision at our recall points
• $P_i(R = r)$ can be measured: the precision at each point when recall changes (because a new relevant document is retrieved)
• Problem: unless the number of relevant documents per query is divisible by 10, $\tilde{P}_i(r_j)$ does not coincide with a measurable data point $r$
• Solution: interpolation

$$\tilde{P}_i(r_j) = \begin{cases} 
\max(r_j \leq r < r_{j+1}) P_i(R = r) & \text{if } P_i(R = r) \text{ exists} \\
\tilde{P}_i(r_{j+1}) & \text{otherwise}
\end{cases}$$

• Note that $P_i(R = 1)$ can always be measured.
11 point average precision: measured data points, Q1

- Blue for Query 1
- Bold Circles measured
- Five $r_f$'s coincide with data-point

11 point average precision: interpolation, Q1

- Blue for Query 1
- Bold Circles measured
- Thin circles interpolated
11 point average precision: measured data points, Q2

- Red for Query 2
- Bold Circles are measured
- Only $r_1(0)$ coincides with a data point

11 point average precision: interpolation, Q2

- Red for Query 2
- Bold Circles measured
- Thin circles interpolated
11 point average precision: averaging

- Now average at each $p_r$
- over $N$ (number of queries)
- $\rightarrow$ 11 data points

11 point average precision: area/result

- End result:
- 11 point average precision
- (representation of area)
11 point average precision: same example as a table

\[
P_1(r_i) \quad \sum_{j=1}^{N} P_j(r_i) \quad P_2(r_i)
\]

<table>
<thead>
<tr>
<th>#</th>
<th>R</th>
<th>(P_1(r_0) = 1.00)</th>
<th>(P_1(r_1) = 1.00)</th>
<th>(P_1(r_2) = P_1(R = .2) = 1.00)</th>
<th>(P_1(r_3) = 0.67)</th>
<th>(P_1(r_4) = P_1(R = .4) = 0.67)</th>
<th>(P_2(r_0) = 1.00)</th>
<th>(P_2(r_1) = 1.00)</th>
<th>(P_2(r_2) = 1.00)</th>
<th>(P_2(r_3) = 1.00)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>0.20</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>(\bar{P}_2(r_0) = 1.00)</td>
<td>(\bar{P}_2(r_1) = 1.00)</td>
<td>(\bar{P}_2(r_2) = 1.00)</td>
<td>(\bar{P}_2(r_3) = 1.00)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>(P_1(r_4) = P_1(R = .6) = 0.50)</td>
<td>(P_2(r_0) = 0.67)</td>
<td>(P_2(r_1) = 0.67)</td>
<td>(P_2(r_2) = 0.67)</td>
<td>(P_2(r_3) = 0.67)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>(P_1(r_5) = 0.50)</td>
<td>(P_2(r_0) = 0.59)</td>
<td>(P_2(r_1) = 0.59)</td>
<td>(P_2(r_2) = 0.59)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>(P_1(r_6) = 0.50)</td>
<td>(P_2(r_0) = 0.50)</td>
<td>(P_2(r_1) = 0.50)</td>
<td>(P_2(r_2) = 0.50)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>(P_1(r_7) = 0.40)</td>
<td>(P_2(r_0) = 0.30)</td>
<td>(P_2(r_1) = 0.30)</td>
<td>(P_2(r_2) = 0.30)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>(P_1(r_8) = P_1(R = .8) = 0.40)</td>
<td>(P_2(r_0) = 0.23)</td>
<td>(P_2(r_1) = 0.23)</td>
<td>(P_2(r_2) = 0.23)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>(P_1(r_9) = 0.25)</td>
<td>(P_2(r_0) = 0.20)</td>
<td>(P_2(r_1) = 0.20)</td>
<td>(P_2(r_2) = 0.20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>(P_1(r_{10}) = P_1(R = 1.0) = 0.25)</td>
<td>(P_2(r_0) = 0.20)</td>
<td>(P_2(r_1) = 0.20)</td>
<td>(P_2(r_2) = 0.20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>(P_1(r_{11}) = P_1(R = .0) = 0.25)</td>
<td>(P_2(r_0) = 0.20)</td>
<td>(P_2(r_1) = 0.20)</td>
<td>(P_2(r_2) = 0.20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>(P_1(r_{12}) = P_1(R = .0) = 0.25)</td>
<td>(P_2(r_0) = 0.20)</td>
<td>(P_2(r_1) = 0.20)</td>
<td>(P_2(r_2) = 0.20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>(P_1(r_{13}) = P_1(R = .0) = 0.25)</td>
<td>(P_2(r_0) = 0.20)</td>
<td>(P_2(r_1) = 0.20)</td>
<td>(P_2(r_2) = 0.20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>(P_1(r_{14}) = P_1(R = .0) = 0.25)</td>
<td>(P_2(r_0) = 0.20)</td>
<td>(P_2(r_1) = 0.20)</td>
<td>(P_2(r_2) = 0.20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>(P_1(r_{15}) = P_1(R = .0) = 0.25)</td>
<td>(P_2(r_0) = 0.20)</td>
<td>(P_2(r_1) = 0.20)</td>
<td>(P_2(r_2) = 0.20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>(P_1(r_{16}) = P_1(R = .0) = 0.25)</td>
<td>(P_2(r_0) = 0.20)</td>
<td>(P_2(r_1) = 0.20)</td>
<td>(P_2(r_2) = 0.20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>(P_1(r_{17}) = P_1(R = .0) = 0.25)</td>
<td>(P_2(r_0) = 0.20)</td>
<td>(P_2(r_1) = 0.20)</td>
<td>(P_2(r_2) = 0.20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>(P_1(r_{18}) = P_1(R = .0) = 0.25)</td>
<td>(P_2(r_0) = 0.20)</td>
<td>(P_2(r_1) = 0.20)</td>
<td>(P_2(r_2) = 0.20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>(P_1(r_{19}) = P_1(R = .0) = 0.25)</td>
<td>(P_2(r_0) = 0.20)</td>
<td>(P_2(r_1) = 0.20)</td>
<td>(P_2(r_2) = 0.20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>(P_1(r_{20}) = P_1(R = .0) = 0.25)</td>
<td>(P_2(r_0) = 0.20)</td>
<td>(P_2(r_1) = 0.20)</td>
<td>(P_2(r_2) = 0.20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>(P_1(r_{21}) = P_1(R = .0) = 0.25)</td>
<td>(P_2(r_0) = 0.20)</td>
<td>(P_2(r_1) = 0.20)</td>
<td>(P_2(r_2) = 0.20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>(P_1(r_{22}) = P_1(R = .0) = 0.25)</td>
<td>(P_2(r_0) = 0.20)</td>
<td>(P_2(r_1) = 0.20)</td>
<td>(P_2(r_2) = 0.20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(\bar{P}_1(r_j)\) is (interpolated) precision of \(i\)th query, at \(j\)th recall point. \(P_i(R = r_j)\) (black) are exactly measured precision values.

Mean Average Precision (MAP)

- Also called “average precision at seen relevant documents”
- Determine precision at each point when a new relevant document gets retrieved
- Use \(P=0\) for each relevant document that was not retrieved
- Determine average for each query, then average over queries

\[
MAP = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{Q_j} \sum_{i=1}^{Q_j} P(doc_i)
\]

with:

- \(Q_j\) number of relevant documents for query \(j\)
- \(N\) number of queries
- \(P(doc_i)\) precision at \(i\)th relevant document
Mean Average Precision: example

MAP favours systems which return relevant documents fast
- Precision-biased

\[ MAP = \frac{0.564 + 0.623}{2} = 0.594 \]

Relevance Judgements and Subjectivity

- Relevance is subjective → Judgements differ across judges
- Relevance is situational → Judgements also differ across time (same judge!)
- Problem: Systems are not comparable if metrics compiled from different judges or at different times will differ
- Countermeasure, Part A: Use guidelines
  - Relevance defined independently of novelty
  - Then, relevance decisions are independent of each other
- Countermeasure, Part B: counteract natural variation by extensive sampling; large populations of users and information needs
- Then: Relative success measurements on systems stable across judges (but not necessarily absolute ones) (Voorhees, 2000)
- Okay if all you want to do is compare systems
• TREC-7 and 8: P(30) between .40 and .45, using long queries and narratives (one team even for short queries); P(10) = .5 even with short queries, > .5 with medium length queries

• Systems must have improved since TREC-4, 5, and 6 → manual performance (sanity check) remained on a plateau of around .6

• The best TREC-8 ad-hoc systems not stat. significantly different → plateau reached? Ad hoc track discontinued after TREC-8.

• New tasks: filtering, web, QA, genomics, interactive, novelty, robust, video, cross-lingual,...

• 2006 is TREC-15. Latest tasks: spam, terabyte, blog, web, legal

### TREC tracks 1992-2002

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad Hoc</td>
<td>18</td>
<td>24</td>
<td>26</td>
<td>23</td>
<td>26</td>
<td>31</td>
<td>42</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Routing</td>
<td>16</td>
<td>25</td>
<td>25</td>
<td>15</td>
<td>16</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interactive</td>
<td>3</td>
<td>11</td>
<td>2</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>4</td>
<td>10</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confusion</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Database Merging</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filtering</td>
<td>4</td>
<td>7</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>15</td>
<td>19</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>9</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLP</td>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech</td>
<td>13</td>
<td>10</td>
<td>10</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross-Language</td>
<td>13</td>
<td>9</td>
<td>13</td>
<td>16</td>
<td>10</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Precision</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Large Corpus</td>
<td>7</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Query</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question Answering</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Web</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Video</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Novelty Detection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genomic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HARD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Robust</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Terabyte</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enterprise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>31</td>
<td>33</td>
<td>36</td>
<td>38</td>
<td>51</td>
<td>56</td>
<td>66</td>
<td>68</td>
<td>87</td>
<td>93</td>
<td>93</td>
<td>103</td>
<td>117</td>
</tr>
</tbody>
</table>
IR evaluation as currently performed (TREC) only covers one small part of the spectrum:
- System performance in batch mode
- Laboratory conditions; not directly involving real users
- Precision and recall measured from large, fixed test collections

However, this evaluation methodology is very stable and mature
- Host of elaborate performance metrics available, e.g. MAP
- Relevance problem solvable (in principle) by query sampling, guidelines, relative system comparisons
- Recall problem solvable (in practice) by pooling methods
- Provable that these methods produce stable evaluation results


Copyrighted – please go to Student Administration for a Copy