Information Retrieval and Web Search

IR Evaluation and IR Standard Text Collections

Instructor: Rada Mihalcea

Some slides in this section are adapted from lectures by Prof. Ray Mooney (UT) and Prof. Razvan Bunescu (Ohio U.)
IR Evaluation Measures

1) How fast does it index?
   - Number of bytes per second.

2) How fast does it search?
   - Latency as a function of queries per second.

3) What is the cost per query?
   - $/query.

4) What is the level of **user happiness**?
   - How can we quantify user happiness?
User Happiness

- Who is the user we are trying to make happy?
  - Web search engine: searcher. Success: Searcher finds what she was looking for. **Measure:** rate of return to this search engine.
  - Web search engine: advertiser. Success: Searcher clicks on ad. **Measure:** clickthrough rate.
  - Ecommerce: buyer. Success: Buyer buys something. **Measures:** time to purchase, fraction of “conversions” of searchers to buyers.
  - Ecommerce: seller. Success: Seller sells something. **Measure:** profit per item sold.
Relevance as Proxy for User Happiness

• User **happiness** ≈ the **relevance** of search results.

• Relevance is assessed relative to the **user need**, *not* the query.
  – Note: **user need** is translated into a **query**.
  – **Information need**: *I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.*
  – **Query**: *red wine white wine heart attack*
  – Assess whether the retrieved document addresses the underlying need, not whether it has these words.
    • Binary Assessments: **Relevant** or **Nonrelevant**.
To measure relevance effectiveness of ad-hoc IR, we need:

1. A document collection.
2. A suite of information needs, expressible as queries.
   - Must be representative of actual user needs.
   - Sample from query logs, if available.
3. Binary assessments of either Relevant or Nonrelevant for each query and each document.
   - Can be more nuanced, e.g., 0, 1, 2, 3, ...
   - Use pooling, when it is unfeasible to assess every \((q, d)\) pair.
Why System Evaluations?

• Is a retrieval system producing the expected results?
• There are many retrieval models/ algorithms/ systems, which one is the best?
• What is the best component for:
  – Ranking function (inner-product, cosine, …)
  – Term selection (stopword removal, stemming…)
  – Term weighting (TF, TF-IDF,…)
• For a fair comparison:
  – Should be all evaluated using the same measures
  – Should be all evaluated on the same collection of documents
  – Should be all evaluated on the same set of questions
Precision and Recall

\[
\text{recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents}}
\]

\[
\text{precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}}
\]
Precision and Recall

- **Precision vs. Recall:**
  - Precision = The ability to retrieve top-ranked documents that are mostly relevant.
  - Recall = The ability of the search to find *all* of the relevant items in the corpus.

- Determining Recall can be difficult

- **Total number of relevant items is sometimes not available** – use **pooling**
  - Sample across the database and perform relevance judgment on these items.
  - Apply different retrieval algorithms to the same database for the same query. The aggregate of relevant items is taken as the total relevant set.
Trade-off between Recall and Precision

Returns relevant documents but misses many useful ones too

The ideal

Returns most relevant documents but also includes lots of irrelevant documents

Precision and Recall are inverse proportional
F-measure

- One measure of performance that takes into account both recall and precision.
- Harmonic mean of recall and precision:

\[ F = \frac{2PR}{P + R} = \frac{2}{\frac{1}{R} + \frac{1}{P}} \]

- Compared to arithmetic mean, both need to be high for harmonic mean to be high.
Parametrized F-measure

- F-measure is an instantiation of the more general $F_\beta$, for $\beta=1$:

$$F_\beta = \frac{(1 + \beta^2)PR}{\beta^2 P + R} = \frac{1 + \beta^2}{\frac{\beta^2}{R} + \frac{1}{P}}$$

- Value of $\beta$ controls trade-off:
  - $\beta = 1$: Equally weight precision and recall (E=F).
  - $\beta > 1$: Weight precision more.
  - $\beta < 1$: Weight recall more.
Ranked Retrieval Measures

• Binary relevance:
  – 11-point Interpolated Precision-Recall Curve
  – R-precision
  – Precision@K (P@K) and Recall@K (R@K)
  – Mean Average Precision (MAP)
  – Mean Reciprocal Rank (MRR)

• Multiple levels of relevance:
  – Normalized Discounted Cumulative Gain (NDCG)
### Recall-Precision Curves

#### An Example

Let total # of relevant docs = 6
Check each new recall point:

<table>
<thead>
<tr>
<th>n</th>
<th>doc #</th>
<th>relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>588</td>
<td>x</td>
</tr>
<tr>
<td>2</td>
<td>589</td>
<td>x</td>
</tr>
<tr>
<td>3</td>
<td>576</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>590</td>
<td>x</td>
</tr>
<tr>
<td>5</td>
<td>986</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>592</td>
<td>x</td>
</tr>
<tr>
<td>7</td>
<td>984</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>988</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>578</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>985</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>103</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>591</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>772</td>
<td>x</td>
</tr>
<tr>
<td>14</td>
<td>990</td>
<td></td>
</tr>
</tbody>
</table>

- **R=1/6=0.167; P=1/1=1**
- **R=2/6=0.333; P=2/2=1**
- **R=3/6=0.5; P=3/4=0.75**
- **R=4/6=0.667; P=4/6=0.667**
- **R=5/6=0.833; P=5/13=0.38**

Missing one relevant document. Never reach 100% recall.
Interpolating a Recall/Precision Curve

- Interpolate a precision value for each standard recall level:
  - \( r_j \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\} \)
  - \( r_0 = 0.0, r_1 = 0.1, \ldots, r_{10} = 1.0 \)

- The interpolated precision at the \( j \)-th standard recall level is the maximum known precision at any recall level between the \( j \)-th and \( (j + 1) \)-th level:

\[
P(r_j) = \max_{r_j \leq r \leq r_{j+1}} P(r)
\]
Interpolating a Recall/Precision Curve: An Example

Precision

Recall
Average Recall/Precision Curve

- Typically average performance over a large set of queries.
- Compute average precision at each standard recall level across all queries.
- Plot average precision/recall curves to evaluate overall system performance on a document/query corpus.

- Average:
  - Micro-average: compute P/R/F once for the entire set of queries
  - Macro-average: average of within-query precision/recall
How To Compare Two or More Systems

- The curve closest to the upper right-hand corner of the graph indicates the best performance.
R-precision

- Precision at the R-th position in the ranking of results for a query that has R relevant documents.

\[
\begin{array}{|c|c|c|}
\hline
n & \text{doc #} & \text{relevant} \\
\hline
1 & 588 & x \\
2 & 589 & x \\
3 & 576 & \\
4 & 590 & x \\
5 & 986 & \\
6 & 592 & x \\
7 & 984 & \\
8 & 988 & \\
9 & 578 & \\
10 & 985 & \\
11 & 103 & \\
12 & 591 & \\
13 & 772 & x \\
14 & 990 & \\
\hline
\end{array}
\]

\[R = \# \text{ of relevant docs} = 6\]

\[\text{R-Precision} = \frac{4}{6} = 0.67\]
1. Set a rank threshold $K$.

2. Compute % of documents relevant in top $K$.
   - Ignores documents ranked lower than $K$.

   - Example:
     - $\text{Prec}@3$ of $2/3$
     - $\text{Prec}@4$ of $2/4$
     - $\text{Prec}@5$ of $3/5$

   - In a similar way we have $\text{Recall}@K$
Mean Average Precision (MAP)

1. Consider rank position of each of the R relevant docs:
   - $K_1, K_2, \ldots K_R$
2. Compute Precision@K for each $K_1, K_2, \ldots K_R$.
3. Average precision = average of $P@K$.

Example: has AvgPrec of $\frac{1}{3} \cdot \left( \frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right) \approx 0.76$

• MAP is Average Precision across multiple queries.
Average Precision

Ranking #1 = (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6) / 6 = 0.78
Ranking #2 = (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6) / 6 = 0.5
Mean Average Precision (MAP)

Average precision query 1 = \( \frac{1.0 + 0.67 + 0.5 + 0.44 + 0.5}{5} = 0.62 \)
Average precision query 2 = \( \frac{0.5 + 0.4 + 0.43}{3} = 0.44 \)
MAP = \( \frac{0.62 + 0.44}{2} = 0.53 \)
Mean Average Precision (MAP)

- If a relevant document never gets retrieved, we assume the precision corresponding to that relevant document to be zero.

- MAP is macro-averaging: each query counts equally.

- A commonly used measure in current IR research, along with P/R/F
Mean Reciprocal Rank

- Consider rank position, $K$, of first relevant doc
  - Could be only clicked doc

- Reciprocal Rank score = $\frac{1}{K}$

- MRR is the mean RR across multiple queries
Multiple Levels of Relevance

- Documents are rarely entirely relevant or non-relevant to a query.

- Many sources of graded relevance judgments:
  - Relevance judgments on a 5-point scale.
  - Averaging among multiple judges.
Cummulative Gain

- With graded relevance judgments, we can compute the gain at each rank.
- **Cumulative Gain** at rank \( n \):
  \[
  CG_n = \sum_{i=1}^{n} rel_i
  \]
  - Where \( rel_i \) is the graded relevance of the document at position \( i \).

<table>
<thead>
<tr>
<th>( n )</th>
<th>doc #</th>
<th>relevance (gain)</th>
<th>( CG_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>588</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>589</td>
<td>0.6</td>
<td>1.6</td>
</tr>
<tr>
<td>3</td>
<td>576</td>
<td>0.0</td>
<td>1.6</td>
</tr>
<tr>
<td>4</td>
<td>590</td>
<td>0.8</td>
<td>2.4</td>
</tr>
<tr>
<td>5</td>
<td>986</td>
<td>0.0</td>
<td>2.4</td>
</tr>
<tr>
<td>6</td>
<td>592</td>
<td>1.0</td>
<td>3.4</td>
</tr>
<tr>
<td>7</td>
<td>984</td>
<td>0.0</td>
<td>3.4</td>
</tr>
<tr>
<td>8</td>
<td>988</td>
<td>0.0</td>
<td>3.4</td>
</tr>
<tr>
<td>9</td>
<td>578</td>
<td>0.0</td>
<td>3.4</td>
</tr>
<tr>
<td>10</td>
<td>985</td>
<td>0.0</td>
<td>3.4</td>
</tr>
<tr>
<td>11</td>
<td>103</td>
<td>0.0</td>
<td>3.4</td>
</tr>
<tr>
<td>12</td>
<td>591</td>
<td>0.0</td>
<td>3.4</td>
</tr>
<tr>
<td>13</td>
<td>772</td>
<td>0.2</td>
<td>3.6</td>
</tr>
<tr>
<td>14</td>
<td>990</td>
<td>0.0</td>
<td>3.6</td>
</tr>
</tbody>
</table>
Discounted Cumulative Gain

- Users care more about high-ranked documents, so we **discount** results by \( \frac{1}{\log_2(rank)} \)

- Popular measures for evaluating web search and related tasks.

- Discounted Cumulative Gain:

\[
DCG_n = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2 i}
\]

<table>
<thead>
<tr>
<th>n</th>
<th>doc #</th>
<th>rel (gain)</th>
<th>CG_n</th>
<th>log_n</th>
<th>DCG_n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>588</td>
<td>1.0</td>
<td>1.0</td>
<td>-</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>589</td>
<td>0.6</td>
<td>1.6</td>
<td>1.00</td>
<td>1.60</td>
</tr>
<tr>
<td>3</td>
<td>576</td>
<td>0.0</td>
<td>1.6</td>
<td>1.58</td>
<td>1.60</td>
</tr>
<tr>
<td>4</td>
<td>590</td>
<td>0.8</td>
<td>2.4</td>
<td>2.00</td>
<td>2.00</td>
</tr>
<tr>
<td>5</td>
<td>986</td>
<td>0.0</td>
<td>2.4</td>
<td>2.32</td>
<td>2.00</td>
</tr>
<tr>
<td>6</td>
<td>592</td>
<td>1.0</td>
<td>3.4</td>
<td>2.58</td>
<td>2.39</td>
</tr>
<tr>
<td>7</td>
<td>984</td>
<td>0.0</td>
<td>3.4</td>
<td>2.81</td>
<td>2.39</td>
</tr>
<tr>
<td>8</td>
<td>988</td>
<td>0.0</td>
<td>3.4</td>
<td>3.00</td>
<td>2.39</td>
</tr>
<tr>
<td>9</td>
<td>578</td>
<td>0.0</td>
<td>3.4</td>
<td>3.17</td>
<td>2.39</td>
</tr>
<tr>
<td>10</td>
<td>985</td>
<td>0.0</td>
<td>3.4</td>
<td>3.32</td>
<td>2.39</td>
</tr>
<tr>
<td>11</td>
<td>103</td>
<td>0.0</td>
<td>3.4</td>
<td>3.46</td>
<td>2.39</td>
</tr>
<tr>
<td>12</td>
<td>591</td>
<td>0.0</td>
<td>3.4</td>
<td>3.58</td>
<td>2.39</td>
</tr>
<tr>
<td>13</td>
<td>772</td>
<td>0.2</td>
<td>3.6</td>
<td>3.70</td>
<td>2.44</td>
</tr>
<tr>
<td>14</td>
<td>990</td>
<td>0.0</td>
<td>3.6</td>
<td>3.81</td>
<td>2.44</td>
</tr>
</tbody>
</table>
### Normalized Discounted Cumulative Gain (NDCG)

- To compare DCGs, normalize values so that an *ideal ranking* would have a **Normalized DCG** of 1.0.
- Ideal ranking:

<table>
<thead>
<tr>
<th>n</th>
<th>doc # (gain)</th>
<th>rel</th>
<th>CG_n</th>
<th>log_n</th>
<th>DCG_n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>588</td>
<td>1.0</td>
<td>1.0</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>589</td>
<td>0.6</td>
<td>1.6</td>
<td>1.00</td>
<td>1.60</td>
</tr>
<tr>
<td>3</td>
<td>576</td>
<td>0.0</td>
<td>1.6</td>
<td>1.58</td>
<td>1.60</td>
</tr>
<tr>
<td>4</td>
<td>590</td>
<td>0.8</td>
<td>2.4</td>
<td>2.00</td>
<td>2.00</td>
</tr>
<tr>
<td>5</td>
<td>986</td>
<td>0.0</td>
<td>2.4</td>
<td>2.32</td>
<td>2.00</td>
</tr>
<tr>
<td>6</td>
<td>592</td>
<td>1.0</td>
<td>3.4</td>
<td>2.58</td>
<td>2.39</td>
</tr>
<tr>
<td>7</td>
<td>984</td>
<td>0.0</td>
<td>3.4</td>
<td>2.81</td>
<td>2.39</td>
</tr>
<tr>
<td>8</td>
<td>988</td>
<td>0.0</td>
<td>3.4</td>
<td>3.00</td>
<td>2.39</td>
</tr>
<tr>
<td>9</td>
<td>578</td>
<td>0.0</td>
<td>3.4</td>
<td>3.17</td>
<td>2.39</td>
</tr>
<tr>
<td>10</td>
<td>985</td>
<td>0.0</td>
<td>3.4</td>
<td>3.32</td>
<td>2.39</td>
</tr>
<tr>
<td>11</td>
<td>103</td>
<td>0.0</td>
<td>3.4</td>
<td>3.46</td>
<td>2.39</td>
</tr>
<tr>
<td>12</td>
<td>591</td>
<td>0.0</td>
<td>3.4</td>
<td>3.58</td>
<td>2.39</td>
</tr>
<tr>
<td>13</td>
<td>772</td>
<td>0.2</td>
<td>3.6</td>
<td>3.70</td>
<td>2.44</td>
</tr>
<tr>
<td>14</td>
<td>990</td>
<td>0.0</td>
<td>3.6</td>
<td>3.81</td>
<td>2.44</td>
</tr>
</tbody>
</table>
Normalized Discounted Cumulative Gain (NDCG)

- Normalize by DCG of the ideal ranking:
  \[ NDCG_n = \frac{DCG_n}{IDCG_n} \]
  - \( NDCG \leq 1 \) at all ranks.

- \( NDCG \) is now comparable across different queries:
  - Useful for contrasting queries with varying numbers of relevant results.
  - Quite popular for Web search.
Evaluation with Clickthrough Data

Strong position bias, so absolute click rates unreliable
Pairwise Relative Ratings

- Pairs of the form: DocA **better than** DocB for a query
  - Doesn’t mean that DocA **relevant** to query

- Now, rather than assess a rank-ordering wrt per-doc relevance assessments

- Assess in terms of conformance with historical pairwise preferences recorded from user clicks

- BUT! Don’t learn and test on the same ranking algorithm
Comparing two rankings to a baseline ranking

- Given a set of pairwise preferences $P$
- We want to measure two rankings $A$ and $B$
- Define a proximity measure between $A$ and $P$
  - And likewise, between $B$ and $P$
- Want to declare the ranking with better proximity to be the winner
- Proximity measure should reward agreements with $P$ and penalize disagreements
Kendall-tau Distance to Compare Rankings

• Generate all pairs for each ranking
• Let $X$ be the number of agreements between a ranking (say $A$) and $P$
• Let $Y$ be the number of disagreements
• Then the Kendall tau distance between $A$ and $P$ is $(X-Y)/(X+Y)$
Exercise

• Assume perfect ranking $P = (1, 2, 3, 4)$
• Assume two candidate rankings $A = (1, 3, 2, 4)$ and $B = (4, 1, 2, 3)$
• Which candidate ranking is closer to the perfect ranking according to Kendall-tau?
A/B Testing at Web Search Engines

- Can exploit an existing user base to provide useful feedback on a **single innovation**

- Randomly send a small fraction (1–10%) of incoming users to a variant of the system that includes a single change.
  - Have most users use the old system

- Judge effectiveness by measuring change in **clickthrough**: the percentage of users that click on the top result (or any result on the first page)

- Probably the evaluation methodology that large search engines trust the most
Amazon Mechanical Turk Testing

- https://requester.mturk.com/
Standard Methodology for Measuring Relevance in IR

• To measure relevance effectiveness of ad-hoc IR, we need:
  1. A document collection.
  2. A suite of information needs, expressible as queries.
     • Must be representative of actual user needs.
     • Sample from query logs, if available.
  3. Binary assessments of either Relevant or Nonrelevant for each query and each document.
     • Can be more nuanced, e.g., 0, 1, 2, 3, ...
     • Use pooling, when it is unfeasible to assess every \((q, d)\) pair.
Early Test Collections

• Previous experiments were based on the SMART collection which is fairly small. (ftp://ftp.cs.cornell.edu/pub/smart)

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Number Of Documents</th>
<th>Number Of Queries</th>
<th>Raw Size (Mbytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CACM</td>
<td>3,204</td>
<td>64</td>
<td>1.5</td>
</tr>
<tr>
<td>CISI</td>
<td>1,460</td>
<td>112</td>
<td>1.3</td>
</tr>
<tr>
<td>CRAN</td>
<td>1,400</td>
<td>225</td>
<td>1.6</td>
</tr>
<tr>
<td>MED</td>
<td>1,033</td>
<td>30</td>
<td>1.1</td>
</tr>
<tr>
<td>TIME</td>
<td>425</td>
<td>83</td>
<td>1.5</td>
</tr>
</tbody>
</table>

• Different researchers used different test collections and evaluation techniques.
The TREC Benchmark

  - Originated from the TIPSTER program sponsored by Defense Advanced Research Projects Agency (DARPA).
  - Became an annual conference in 1992, co-sponsored by the National Institute of Standards and Technology (NIST) and DARPA.
  - Participants are given parts of a standard set of documents and TOPICS (from which queries have to be derived) in different stages for training and testing.
  - Participants submit the P/R values for the final document and query corpus and present their results at the conference.
TREC Objectives

• Provide a common ground for comparing different IR techniques.
  – Same set of documents and queries, and same evaluation method

• Sharing of resources and experiences in developing the benchmark.
  – With major sponsorship from government to develop large benchmark collections

• Encourage participation from industry and academia.

• Development of new evaluation techniques, particularly for new applications.
  – Retrieval, routing/filtering, non-English collection, web-based collection, question answering
TREC Advantages

• Large scale (compared to a few MB in the SMART Collection).

• Relevance judgments provided.

• Under continuous development with support from the U.S. Government.

• Wide participation:
  – TREC 1: 28 papers 360 pages.
  – TREC 4: 37 papers 560 pages.
  – TREC 7: 61 papers 600 pages.
  – TREC 8: 74 papers.
TREC Tasks

• **Ad hoc**: New questions are being asked on a static set of data.

• **Routing**: Same questions are being asked, but new information is being searched. (news clipping, library profiling).

• New tasks added after TREC 5:
  – Interactive, multilingual, natural language, multiple database merging, filtering, very large corpus (20 GB, 7.5 million documents), question answering.
The TREC Collection

- Both long and short documents (from a few hundred to over one thousand unique terms in a document).
  - Both SGML documents and SGML queries contain many different kinds of information (fields).
  - Generation of the formal queries (Boolean, Vector Space, etc.) is the responsibility of the system.
    - A system may be very good at ranking, but if it generates poor queries from the topic, its final P/R would be poor.

- Test documents consist of:

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>Associate Press Newswire (1989)</td>
<td>514 M</td>
</tr>
<tr>
<td>ZIFF</td>
<td>Computer Select Disks (Ziff-Davis Publishing)</td>
<td>493 M</td>
</tr>
<tr>
<td>FR</td>
<td>Federal Register</td>
<td>469 M</td>
</tr>
<tr>
<td>DOE</td>
<td>Abstracts from Department of Energy reports</td>
<td>190 M</td>
</tr>
</tbody>
</table>
John Blair & Co. is close to an agreement to sell its TV station advertising representation operation and program production unit to an investor group led by James H. Rosenfield, a former CBS Inc. executive, industry sources said. Industry sources put the value of the proposed acquisition at more than $100 million. ...
Sample SGML Query

<head> Tipster Topic Description
<num> Number: 066
<dom> Domain: Science and Technology
<title> Topic: Natural Language Processing
<desc> Description: Document will identify a type of natural language processing technology which is being developed or marketed in the U.S.
<narr> Narrative: A relevant document will identify a company or institution developing or marketing a natural language processing technology, identify the technology, and identify one of more features of the company's product.
<con> Concept(s): 1. natural language processing ;2. translation, language, dictionary
<fac> Factor(s):
<nat> Nationality: U.S.</nat>
</fac>
<def> Definitions(s):
</top>
TREC Evaluation

- **Summary table statistics**: Number of topics, number of documents retrieved, number of relevant documents.

- **11-point Interpolated Precision**: Average precision at 11 recall levels (0 to 1 at 0.1 increments).

- **Document level average**: Average precision when 5, 10, .., 100, ... 1000 documents are retrieved.

- **Average precision histogram**: Difference of the R-precision for each topic and the average R-precision of all systems for that topic.