Evaluation

INST 734
Lecture 5

February 26, 2014
Outline

Recap

Introduction

Unranked evaluation

Ranked evaluation

Benchmarks
Indexing

- Index structure
Indexing

- Index structure
- Index construction
Indexing

- Index structure
- Index construction
- Searching a term
Ranking

- How frequent are the query words in a document?
Ranking

- How frequent are the query words in a document?
  - Term importance: TF.IDF measure
How frequent are the query words in a document?

- Term importance: TF.IDF measure
- Vector space model
How frequent are the query words in a document?

- Term importance: TF.IDF measure
- Vector space model
- Language model
Interaction

- Search interface
Interaction

- Search interface
- Query formulation
Interaction

- Search interface
- Query formulation
- Query modification
Interaction

- Search interface
- Query formulation
- Query modification
- Result examination
Three Major Criteria

- Ease of use
Three Major Criteria

- Ease of use
- Efficiency
Three Major Criteria

- Ease of use
- Efficiency
- Result relevance (most important!)
Take-away today
Take-away today

- Introduction to evaluation: Measures of an IR system
Take-away today

- Introduction to evaluation: Measures of an IR system
- Evaluation of unranked and ranked retrieval
Take-away today

- Introduction to evaluation: Measures of an IR system
- Evaluation of unranked and ranked retrieval
- Evaluation benchmarks
Outline

Recap

Introduction

Unranked evaluation

Ranked evaluation

Benchmarks
Measures for a search engine
Measures for a search engine

- How fast does it index?
Measures for a search engine

- How fast does it index?
  - e.g., number of bytes per hour
Measures for a search engine

- How fast does it index?
  - e.g., number of bytes per hour
- How fast does it search?
Measures for a search engine

- How fast does it index?
  - e.g., number of bytes per hour
- How fast does it search?
  - e.g., latency as a function of queries per second
Measures for a search engine

- How fast does it index?
  - e.g., number of bytes per hour
- How fast does it search?
  - e.g., latency as a function of queries per second
- What is the cost per query?
Measures for a search engine

- How fast does it index?
  - e.g., number of bytes per hour
- How fast does it search?
  - e.g., latency as a function of queries per second
- What is the cost per query?
  - in dollars
Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money
Measures for a search engine

- All of the preceding criteria are *measurable*: we can quantify speed / size / money
- However, the key measure for a search engine is *user happiness*. 
Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money
- However, the key measure for a search engine is user happiness.
- What is user happiness?
Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money
- However, the key measure for a search engine is user happiness.
- What is user happiness?
- Factors include:
Measures for a search engine

- All of the preceding criteria are **measurable**: we can quantify speed / size / money
- However, the key measure for a search engine is **user happiness**.
- What is user happiness?
- Factors include:
  - Speed of response
Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money
- However, the key measure for a search engine is user happiness.
- What is user happiness?
- Factors include:
  - Speed of response
  - Uncluttered User interface
Measures for a search engine

- All of the preceding criteria are **measurable**: we can quantify speed / size / money
- However, the key measure for a search engine is **user happiness**.
- What is user happiness?
- Factors include:
  - Speed of response
  - Uncluttered User interface
  - Most important: **relevance**
Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money
- However, the key measure for a search engine is user happiness.
- What is user happiness?
- Factors include:
  - Speed of response
  - Uncluttered User interface
  - Most important: relevance
- Note that none of these is sufficient: very fast, but useless answers won’t make a user happy.
Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money.
- However, the key measure for a search engine is user happiness.
- What is user happiness?
- Factors include:
  - Speed of response
  - Uncluttered User interface
  - Most important: relevance
- Note that none of these is sufficient: very fast, but useless answers won’t make a user happy.
- How can we quantify user happiness?
Who is the user?

- Who is the user we are trying to make happy?
Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Success: Searcher finds what she was looking for.
Who is the user we are trying to make happy?

Web search engine: searcher. Success: Searcher finds what she was looking for.

Web search engine: advertiser. Success: Searcher clicks on ad.
Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Success: Searcher finds what she was looking for.
- Web search engine: advertiser. Success: Searcher clicks on ad.
Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Success: Searcher finds what she was looking for.
- Web search engine: advertiser. Success: Searcher clicks on ad.
Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Success: Searcher finds what she was looking for.
- Web search engine: advertiser. Success: Searcher clicks on ad.
- Enterprise: CEO. Success: Employees are more productive (because of effective search).
Who is the user?

- 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.


- Enterprise: CEO. Success: Employees are more productive (because of effective search).
Who is the user?

- 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**
- Enterprise: CEO. Success: Employees are more productive (because of effective search).
Who is the user?

- 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
- E-commerce: buyer. Success: Buyer buys something. **Measures:** time to purchase, fraction of “conversions” of searchers to buyers

- Enterprise: CEO. Success: Employees are more productive (because of effective search).
Who is the user?

- 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
- E-commerce: buyer. Success: Buyer buys something. **Measures:** time to purchase, fraction of “conversions” of searchers to buyers
- E-commerce: seller. Success: Seller sells something. **Measure:** profit per item sold
- Enterprise: CEO. Success: Employees are more productive (because of effective search).
Who is the user?

- 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
- E-commerce: buyer. Success: Buyer buys something. **Measures**: time to purchase, fraction of “conversions” of searchers to buyers
- E-commerce: seller. Success: Seller sells something. **Measure**: profit per item sold
- Enterprise: CEO. Success: Employees are more productive (because of effective search). **Measure**: profit of the company
Most common definition of user happiness: Relevance
Most common definition of user happiness: Relevance

- User happiness is equated with the relevance of search results to the query.
Most common definition of user happiness: Relevance

- User happiness is equated with the relevance of search results to the query.
- But how do you measure relevance?
Most common definition of user happiness: Relevance

- User happiness is equated with the relevance of search results to the query.
- But how do you measure relevance?
- Standard methodology in information retrieval consists of three elements.
User happiness is equated with the relevance of search results to the query.

But how do you measure relevance?

Standard methodology in information retrieval consists of three elements.

- A benchmark document collection
Most common definition of user happiness: Relevance

- User happiness is equated with the relevance of search results to the query.
- But how do you measure relevance?
- Standard methodology in information retrieval consists of three elements.
  - A benchmark document collection
  - A benchmark suite of queries
Most common definition of user happiness: Relevance

- User happiness is equated with the relevance of search results to the query.
- But how do you measure relevance?
- Standard methodology in information retrieval consists of three elements.
  - A benchmark document collection
  - A benchmark suite of queries
  - An assessment of the relevance of each query-document pair
Relevance: query vs. information need
Relevance: query vs. information need

- Relevance to what?
Relevance: query vs. information need

- Relevance to what?
- First take: relevance to the query
Relevance: query vs. information need

- Relevance to what?
- First take: relevance to the query
- “Relevance to the query” is very problematic.
Relevance: query vs. information need

- Relevance to what?
- First take: relevance to the query
- “Relevance to the query” is very problematic.
- Information need $i$: “I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.”
Relevance: query vs. information need

- Relevance to what?
- First take: relevance to the query
  - “Relevance to the query” is very problematic.
- Information need $i$: “I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.”
- This is an information need, not a query.
Relevance: query vs. information need

- Relevance to what?
- First take: relevance to the query
- “Relevance to the query” is very problematic.
- Information need $i$: “I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.”
- This is an information need, not a query.
- Query $q$: [red wine white wine heart attack]
Relevance: query vs. information need

- Relevance to what?
- First take: relevance to the query
- “Relevance to the query” is very problematic.
- Information need $i$: “I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.”
- This is an information need, not a query.
- Query $q$: [red wine white wine heart attack]
- Consider document $d'$: At the heart of his speech was an attack on the wine industry lobby for downplaying the role of red and white wine in drunk driving.
Relevance: query vs. information need

- Relevance to what?
- First take: relevance to the query
- “Relevance to the query” is very problematic.
- **Information need** \( i \): “I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.”
- This is an information need, not a query.
- **Query** \( q \): [red wine white wine heart attack]
- Consider document \( d' \): *At the heart of his speech was an attack on the wine industry lobby for downplaying the role of red and white wine in drunk driving.*
- \( d' \) is an excellent match for query \( q \) ...
Relevance: query vs. information need

- Relevance to what?
- First take: relevance to the query
- “Relevance to the query” is very problematic.
- Information need $i$: “I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.”
- This is an information need, not a query.
- Query $q$: [red wine white wine heart attack]
- Consider document $d'$: At the heart of his speech was an attack on the wine industry lobby for downplaying the role of red and white wine in drunk driving.
- $d'$ is an excellent match for query $q$ . . .
- $d'$ is not relevant to the information need $i$. 
Relevance: query vs. information need
User happiness can only be measured by relevance to an information need, not by relevance to queries.
User happiness can only be measured by relevance to an information need, not by relevance to queries.

Our terminology is sloppy in these slides: we talk about query-document relevance judgments even though we mean information-need-document relevance judgments.
Outline

Recap

Introduction

Unranked evaluation

Ranked evaluation

Benchmarks
Precision and recall

- Precision \((P)\) is the fraction of retrieved documents that are relevant.

\[
\text{Precision} = \frac{\# \text{(relevant items retrieved)}}{\# \text{(retrieved items)}} = P(\text{relevant} | \text{retrieved})
\]
Precision and recall

- Precision ($P$) is the fraction of retrieved documents that are relevant
  \[
  \text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})
  \]

- Recall ($R$) is the fraction of relevant documents that are retrieved
  \[
  \text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})
  \]
Precision and recall
Precision and recall

<table>
<thead>
<tr>
<th></th>
<th>Relevant</th>
<th>Nonrelevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved</td>
<td>true positives (TP)</td>
<td>false positives (FP)</td>
</tr>
<tr>
<td>Not retrieved</td>
<td>false negatives (FN)</td>
<td>true negatives (TN)</td>
</tr>
</tbody>
</table>

\[
P = \frac{TP}{TP + FP} \\
R = \frac{TP}{TP + FN}
\]
Precision/recall tradeoff
Precision/recall tradeoff

- You can increase recall by returning more docs.
Precision/recall tradeoff

- You can increase recall by returning more docs.
- Recall is a non-decreasing function of the number of docs retrieved.
You can increase recall by returning more docs.
Recall is a non-decreasing function of the number of docs retrieved.
A system that returns all docs has 100% recall!
Precision/recall tradeoff

- You can increase recall by returning more docs.
- Recall is a non-decreasing function of the number of docs retrieved.
- A system that returns all docs has 100% recall!
- The converse is also true (usually): It's easy to get high precision for very low recall.
You can increase recall by returning more docs.
Recall is a non-decreasing function of the number of docs retrieved.
A system that returns all docs has 100% recall!
The converse is also true (usually): It’s easy to get high precision for very low recall.
Suppose the document with the largest score is relevant. How can we maximize precision?
A combined measure: $F$
A combined measure: $F$

- $F$ allows us to trade off precision against recall.
A combined measure: $F$

- $F$ allows us to trade off precision against recall.

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}$$
A combined measure: $F$

- $F$ allows us to trade off precision against recall.
- $F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}$
- $\alpha \in [0, 1]$
A combined measure: $F$

- $F$ allows us to trade off precision against recall.

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}$$

- $\alpha \in [0, 1]$
- Most frequently used: balanced $F$ with $\alpha = 0.5$
A combined measure: $F$

- $F$ allows us to trade off precision against recall.

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

- $\alpha \in [0, 1]$
- Most frequently used: balanced $F$ with $\alpha = 0.5$
  - This is the harmonic mean of $P$ and $R$: $\frac{1}{F} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$
A combined measure: $F$

- $F$ allows us to trade off precision against recall.

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}$$

- $\alpha \in [0, 1]$
- Most frequently used: balanced $F$ with $\alpha = 0.5$
  - This is the harmonic mean of $P$ and $R$: $\frac{1}{F} = \frac{1}{2} \left( \frac{1}{P} + \frac{1}{R} \right)$
- What value range of $\alpha$ weights recall higher than precision?
Example for precision, recall, F1
Example for precision, recall, F1

<table>
<thead>
<tr>
<th></th>
<th>relevant</th>
<th>not relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>not retrieved</td>
<td>60</td>
<td>1,000,000</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>1,000,040</td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td>60</td>
<td>1,000,060</td>
</tr>
<tr>
<td>not retrieved</td>
<td>1,000,040</td>
<td>1,000,120</td>
</tr>
</tbody>
</table>
Example for precision, recall, F1

<table>
<thead>
<tr>
<th></th>
<th>relevant</th>
<th>not relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>not retrieved</td>
<td>60</td>
<td>1,000,000</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>1,000,040</td>
</tr>
</tbody>
</table>
Example for precision, recall, F1

<table>
<thead>
<tr>
<th></th>
<th>relevant</th>
<th>not relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>not retrieved</td>
<td>60</td>
<td>1,000,000</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>1,000,040</td>
</tr>
</tbody>
</table>

\[ P = \frac{20}{20 + 40} = \frac{1}{3} \]
Example for precision, recall, F1

<table>
<thead>
<tr>
<th></th>
<th>relevant</th>
<th>not relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>not retrieved</td>
<td>60</td>
<td>1,000,000</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>1,000,040</td>
</tr>
</tbody>
</table>

- \( P = \frac{20}{20 + 40} = \frac{1}{3} \)
- \( R = \frac{20}{20 + 60} = \frac{1}{4} \)
Example for precision, recall, F1

<table>
<thead>
<tr>
<th></th>
<th>relevant</th>
<th>not relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>not retrieved</td>
<td>60</td>
<td>1,000,000</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>1,000,040</td>
</tr>
</tbody>
</table>

- \( P = \frac{20}{20 + 40} = \frac{1}{3} \)
- \( R = \frac{20}{20 + 60} = \frac{1}{4} \)
- \( F_1 = 2 \cdot \frac{\frac{1}{3} + \frac{1}{4}}{\frac{1}{3} + \frac{1}{4} + \frac{1}{4}} = \frac{2}{7} \)
Accuracy
 Accuracy

- Why do we use measures like precision, recall, and $F$?
Accuracy

- Why do we use measures like precision, recall, and $F$?
- Why not something simple like accuracy?
Accuracy

- Why do we use measures like precision, recall, and $F$?
- Why not something simple like accuracy?
- Accuracy is the fraction of decisions (relevant/nonrelevant) that are correct.
Accuracy

- Why do we use measures like precision, recall, and $F$?
- Why not something simple like accuracy?
- Accuracy is the fraction of decisions (relevant/nonrelevant) that are correct.
- In terms of the contingency table above, 
  accuracy $= \frac{TP + TN}{TP + FP + FN + TN}$. 
Exercise
Exercise

- Compute precision, recall and $F_1$ for this result set:

<table>
<thead>
<tr>
<th></th>
<th>relevant</th>
<th>not relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>not retrieved</td>
<td>82</td>
<td>1,000,000,000</td>
</tr>
</tbody>
</table>
Exercise

- Compute precision, recall and $F_1$ for this result set:
  
<table>
<thead>
<tr>
<th></th>
<th>relevant</th>
<th>not relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>not retrieved</td>
<td>82</td>
<td>1,000,000,000</td>
</tr>
</tbody>
</table>

- The snoogle search engine below always returns 0 results (“0 matching results found”), regardless of the query. Why does snoogle demonstrate that accuracy is not a useful measure in IR?
Why accuracy is a useless measure in IR
Why accuracy is a useless measure in IR

- Simple trick to maximize accuracy in IR: always say no and return nothing
Why accuracy is a useless measure in IR

- Simple trick to maximize accuracy in IR: always say no and return nothing
- You then get 99.99% accuracy on most queries.
Why accuracy is a useless measure in IR

- Simple trick to maximize accuracy in IR: always say no and return nothing.
- You then get 99.99% accuracy on most queries.
- Searchers on the web (and in IR in general) want to find something and have a certain tolerance for junk.
Why accuracy is a useless measure in IR

- Simple trick to maximize accuracy in IR: always say no and return nothing
- You then get 99.99% accuracy on most queries.
- Searchers on the web (and in IR in general) want to find something and have a certain tolerance for junk.
- It’s better to return some bad hits as long as you return something.
Why accuracy is a useless measure in IR

- Simple trick to maximize accuracy in IR: always say no and return nothing
- You then get 99.99% accuracy on most queries.
- Searchers on the web (and in IR in general) want to find something and have a certain tolerance for junk.
- It’s better to return some bad hits as long as you return something.
- → We use precision, recall, and $F$ for evaluation, not accuracy.
F: Why harmonic mean?
F: Why harmonic mean?

- Why don’t we use a different mean of $P$ and $R$ as a measure?
F: Why harmonic mean?

- Why don’t we use a different mean of $P$ and $R$ as a measure?
  - e.g., the arithmetic mean
F: Why harmonic mean?

- Why don’t we use a different mean of $P$ and $R$ as a measure?
  - e.g., the arithmetic mean
- The simple (arithmetic) mean is 50% for “return-everything” search engine, which is too high.
F: Why harmonic mean?

- Why don’t we use a different mean of $P$ and $R$ as a measure?
  - e.g., the arithmetic mean
- The simple (arithmetic) mean is 50% for “return-everything” search engine, which is too high.
- Desideratum: Punish really bad performance on either precision or recall.
F: Why harmonic mean?

- Why don’t we use a different mean of $P$ and $R$ as a measure?
  - e.g., the arithmetic mean
- The simple (arithmetic) mean is 50% for “return-everything” search engine, which is too high.
- Desideratum: Punish really bad performance on either precision or recall.
- Taking the minimum achieves this.
F: Why harmonic mean?

- Why don’t we use a different mean of $P$ and $R$ as a measure?
  - e.g., the arithmetic mean
- The simple (arithmetic) mean is 50% for “return-everything” search engine, which is too high.
- Desideratum: Punish really bad performance on either precision or recall.
- Taking the minimum achieves this.
- But minimum is not smooth and hard to weight.
Why harmonic mean?

- Why don’t we use a different mean of $P$ and $R$ as a measure?
  - e.g., the arithmetic mean
- The simple (arithmetic) mean is 50% for “return-everything” search engine, which is too high.
- Desideratum: Punish really bad performance on either precision or recall.
- Taking the minimum achieves this.
- But minimum is not smooth and hard to weight.
- $F$ (harmonic mean) is a kind of smooth minimum.
Difficulties in using precision, recall and $F$
Difficulties in using precision, recall and $F$

- We need relevance judgments for information-need-document pairs – but they are expensive to produce.
Difficulties in using precision, recall and $F$

- We need relevance judgments for information-need-document pairs – but they are expensive to produce.
- For alternatives to using precision/recall and having to produce relevance judgments – see end of this lecture.
Outline

Recap

Introduction

Unranked evaluation

Ranked evaluation

Benchmarks
Rank Based Measures

- Binary Relevance (good or bad)
Rank Based Measures

- Binary Relevance (good or bad)
  1. Precision@k
Rank Based Measures

- Binary Relevance (good or bad)
  1. Precision@k
  2. Mean average precision (MAP)
Rank Based Measures

- Binary Relevance (good or bad)
  1. Precision@k
  2. Mean average precision (MAP)
  3. Mean Reciprocal Rank (MRR)
Rank Based Measures

- Binary Relevance (good or bad)
  1. Precision@k
  2. Mean average precision (MAP)
  3. Mean Reciprocal Rank (MRR)

- Multiple levels of relevance (excellent, very good, good, bad)
Rank Based Measures

- Binary Relevance (good or bad)
  1. Precision@k
  2. Mean average precision (MAP)
  3. Mean Reciprocal Rank (MRR)

- Multiple levels of relevance (excellent, very good, good, bad)
  1. Discounted Cumulative Gain (DCG)
Rank Based Measures

- Binary Relevance (good or bad)
  1. Precision@k
  2. Mean average precision (MAP)
  3. Mean Reciprocal Rank (MRR)

- Multiple levels of relevance (excellent, very good, good, bad)
  1. Discounted Cumulative Gain (DCG)
  2. Normalized Discounted Cumulative Gain (NDCG)
Precision@K

- Set a rank position K
Precision@K

- Set a rank position $K$
- Compute % relevant in top $K$
Precision@K

- Set a rank position K
- Compute % relevant in top K
- Ignore documents ranked lower than K
Precision@K

- Set a rank position K
- Compute % relevant in top K
- Ignore documents ranked lower than K
- Example ranked list: $D_1 \ D_2 \ D_3 \ D_4 \ D_5$
Set a rank position $K$
Compute % relevant in top $K$
Ignore documents ranked lower than $K$
Example ranked list: $D_1 \ D_2 \ D_3 \ D_4 \ D_5$

Prec@3 = 2/3
Precision@K

- Set a rank position $K$
- Compute % relevant in top $K$
- Ignore documents ranked lower than $K$
- Example ranked list: $D_1 \ D_2 \ D_3 \ D_4 \ D_5$
  - $\text{Prec@3} = \frac{2}{3}$
  - $\text{Prec@4} = \frac{2}{4}$
Precision@K

- Set a rank position $K$
- Compute % relevant in top $K$
- Ignore documents ranked lower than $K$
- Example ranked list: $D_1\ D_2\ D_3\ D_4\ D_5$
  - Prec@3 = 2/3
  - Prec@4 = 2/4
  - Prec@5 = 3/5
Precision@K

- Set a rank position $K$
- Compute % relevant in top $K$
- Ignore documents ranked lower than $K$
- Example ranked list: $D_1 \ D_2 \ D_3 \ D_4 \ D_5$
  - $\text{Prec@3} = \frac{2}{3}$
  - $\text{Prec@4} = \frac{2}{4}$
  - $\text{Prec@5} = \frac{3}{5}$
- In similar fashion we can compute Recall@K
Mean Average Precision

- Consider rank position of each relevant doc
Mean Average Precision

- Consider rank position of each relevant doc
- Compute precision at each relevant document
Mean Average Precision

- Consider rank position of each relevant doc
- Compute precision at each relevant document
- Average precision = average of P@K
Mean Average Precision

- Consider rank position of each relevant doc
- Compute precision at each relevant document
- Average precision = average of P@K
- Example ranked list: $D_1 \ D_2 \ D_3 \ D_4 \ D_5$
Mean Average Precision

- Consider rank position of each relevant doc
- Compute precision at each relevant document
- Average precision = average of $P@K$
- Example ranked list: $D_1\ D_2\ D_3\ D_4\ D_5$
  - Average precision: $\frac{1}{3} \times \left( \frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right)$
Mean Average Precision

- Consider rank position of each relevant doc
- Compute precision at each relevant document
- Average precision = average of P@K
- Example ranked list: $D_1 \ D_2 \ D_3 \ D_4 \ D_5$
  - Average precision: $\frac{1}{3} \times (\frac{1}{1} + \frac{2}{3} + \frac{3}{5})$
- MAP is Average Precision across multiple queries/rankings
Average Precision: Example

= the relevant documents

<table>
<thead>
<tr>
<th>Ranking #1</th>
<th>Recall</th>
<th>0.17 0.17 0.33 0.50 0.67 0.83 0.83 0.83 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>1.0 0.5 0.67 0.75 0.80 0.83 0.71 0.63 0.56 0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ranking #2</th>
<th>Recall</th>
<th>0.0 0.17 0.17 0.17 0.33 0.50 0.67 0.83 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>0.0 0.5 0.33 0.25 0.40 0.50 0.57 0.5 0.56 0.6</td>
</tr>
</tbody>
</table>

Ranking #1: \[
\frac{(1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)}{6} = 0.78
\]

Ranking #2: \[
\frac{(0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)}{6} = 0.52
\]
Mean Average Precision: Example

average precision query 1 = \( \frac{1.0 + 0.67 + 0.5 + 0.44 + 0.5}{5} = 0.62 \)

average precision query 2 = \( \frac{0.5 + 0.4 + 0.43}{3} = 0.44 \)

mean average precision = \( \frac{0.62 + 0.44}{2} = 0.53 \)
Mean average precision

- Good for web search?
Mean average precision

- Good for web search?
  - MAP assumes user is interested in finding many relevant documents for each query
Mean average precision

- Good for web search?
  - MAP assumes user is interested in finding many relevant documents for each query
  - MAP requires many relevance judgments in text collection
Mean average precision

- Good for web search?
  - MAP assumes user is interested in finding many relevant documents for each query
  - MAP requires many relevance judgments in text collection
  - Does not differentiate excellent one from good one
Beyond Binary relevance
Discounted Cumulative Gain

- Popular measure for evaluating web search and related tasks
Discounted Cumulative Gain

- Popular measure for evaluating web search and related tasks
- Two assumptions:
Discounted Cumulative Gain

- Popular measure for evaluating web search and related tasks
- Two assumptions:
  - Highly relevant documents are more useful than marginally relevant documents
Discounted Cumulative Gain

- Popular measure for evaluating web search and related tasks
- Two assumptions:
  - Highly relevant documents are more useful than marginally relevant documents
  - The lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined
Discounted Cumulative Gain

- Uses graded relevance as a measure of usefulness, or gain, from examining a document
Discounted Cumulative Gain

- Uses graded relevance as a measure of usefulness, or gain, from examining a document
- Gain is accumulated starting at the top of the ranking and may be reduced, or discounted, at lower ranks
Discounted Cumulative Gain

- Uses graded relevance as a measure of usefulness, or gain, from examining a document
- Gain is accumulated starting at the top of the ranking and may be reduced, or discounted, at lower ranks
- Typical discount is $\frac{1}{\log(rank)}$
Discounted Cumulative Gain

- Uses graded relevance as a measure of usefulness, or gain, from examining a document
- Gain is accumulated starting at the top of the ranking and may be reduced, or discounted, at lower ranks
- Typical discount is \( \frac{1}{\log(rank)} \)
  - With base 2, the discount at rank 4 is 1/2, and at rank 8 it is 1/3
What if relevance judgments are in a scale of $[0, r]$? $r > 2$
Summarize a Ranking: DCG

- What if relevance judgments are in a scale of $[0, r]$? $r > 2$
- Cumulative Gain (CG) at rank $n$
Summarize a Ranking: DCG

- What if relevance judgments are in a scale of $[0, r]$? $r > 2$
- Cumulative Gain (CG) at rank n
  - Let the ratings of the n documents be $r_1, r_2, \ldots, r_n$ (in ranked order)
What if relevance judgments are in a scale of $[0, r]$? $r > 2$

Cumulative Gain (CG) at rank $n$

- Let the ratings of the $n$ documents be $r_1 \ r_2 \ \ldots \ \ r_n$ (in ranked order)
- $CG = r_1 + r_2 + r_n$
What if relevance judgments are in a scale of \([0, r]\)? \(r > 2\)

- **Cumulative Gain (CG) at rank n**
  - Let the ratings of the n documents be \(r_1 \ r_2 \ \ldots \ r_n\) (in ranked order)
  - \(CG = r_1 + r_2 + r_n\)

- **Discounted Cumulative Gain (DCG) at rank n**
What if relevance judgments are in a scale of $[0, r]$? $r > 2$

Cumulative Gain (CG) at rank $n$
- Let the ratings of the $n$ documents be $r_1, r_2, \ldots, r_n$ (in ranked order)
- $CG = r_1 + r_2 + r_n$

Discounted Cumulative Gain (DCG) at rank $n$
- $DCG = r_1 + \frac{r_2}{\log_2(2)} + \frac{r_3}{\log_2(3)} + \cdots + \frac{r_n}{\log_2(n)}$
What if relevance judgments are in a scale of \([0, r]\)? \(r > 2\)

Cumulative Gain (CG) at rank \(n\)
- Let the ratings of the \(n\) documents be \(r_1 \ r_2 \ \ldots \ r_n\) (in ranked order)
- \(CG = r_1 + r_2 + r_n\)

Discounted Cumulative Gain (DCG) at rank \(n\)
- \(DCG = r_1 + \frac{r_2}{\log_2(2)} + \frac{r_3}{\log_2(3)} + \ldots + \frac{r_n}{\log_2(n)}\)
- We may use any base for the logarithm
DCG: Example

- 10 ranked documents judged on 0-3 relevance scale:
  3, 2, 3, 0, 0, 1, 2, 2, 3, 0
DCG: Example

- 10 ranked documents judged on 0-3 relevance scale:
  3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- discounted gain:
  3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0
  $= 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0$
DCG: Example

- 10 ranked documents judged on 0-3 relevance scale:
  3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- discounted gain:
  3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0
  = 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0
- DCG:
  3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61
Summarize a Ranking: NDCG

- Normalized Discounted Cumulative Gain (NDCG) at rank n
Summarize a Ranking: NDCG

- Normalized Discounted Cumulative Gain (NDCG) at rank n
  - Normalize DCG at rank n by the DCG value at rank n of the ideal rankings
Summarize a Ranking: NDCG

- Normalized Discounted Cumulative Gain (NDCG) at rank $n$
  - Normalize DCG at rank $n$ by the DCG value at rank $n$ of the ideal rankings
  - The ideal ranking would first return the documents with the highest relevance level, then the next highest relevance level, etc
Summarize a Ranking: NDCG

- Normalized Discounted Cumulative Gain (NDCG) at rank $n$
  - Normalize DCG at rank $n$ by the DCG value at rank $n$ of the ideal rankings
  - The ideal ranking would first return the documents with the highest relevance level, then the next highest relevance level, etc
- Why normalization?
Summarize a Ranking: NDCG

- Normalized Discounted Cumulative Gain (NDCG) at rank n
  - Normalize DCG at rank n by the DCG value at rank n of the ideal rankings
  - The ideal ranking would first return the documents with the highest relevance level, then the next highest relevance level, etc

- Why normalization?
  - Normalization useful for contrasting queries with varying numbers of relevant results
Summarize a Ranking: NDCG

- Normalized Discounted Cumulative Gain (NDCG) at rank n
  - Normalize DCG at rank n by the DCG value at rank n of the ideal rankings
  - The ideal ranking would first return the documents with the highest relevance level, then the next highest relevance level, etc

- Why normalization?
  - Normalization useful for contrasting queries with varying numbers of relevant results

- NDCG is now quite popular in evaluating Web search
### NDCG: Example

#### 4 documents: \(d_1, d_2, d_3, d_4\)

<table>
<thead>
<tr>
<th>(i)</th>
<th>Ground Truth</th>
<th>(r_i)</th>
<th>Ranking Function(_1)</th>
<th>(r_i)</th>
<th>Ranking Function(_2)</th>
<th>(r_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(d_4)</td>
<td>2</td>
<td>(d_3)</td>
<td>2</td>
<td>(d_3)</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>(d_3)</td>
<td>2</td>
<td>(d_4)</td>
<td>2</td>
<td>(d_2)</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>(d_2)</td>
<td>1</td>
<td>(d_2)</td>
<td>1</td>
<td>(d_4)</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>(d_1)</td>
<td>0</td>
<td>(d_1)</td>
<td>0</td>
<td>(d_1)</td>
<td>0</td>
</tr>
</tbody>
</table>

\(\text{NDCG}_{\text{GT}} = 1.00\)
\(\text{NDCG}_{\text{RF}_1} = 1.00\)
\(\text{NDCG}_{\text{RF}_2} = 0.9203\)

\[
\text{DCG}_{\text{GT}} = 2 + \left( \frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309
\]

\[
\text{DCG}_{\text{RF}_1} = 2 + \left( \frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309
\]

\[
\text{DCG}_{\text{RF}_2} = 2 + \left( \frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.2619
\]

\[
\text{MaxDCG} = \text{DCG}_{\text{GT}} = 4.6309
\]
What if the results are not in a list?

- Suppose there’s only one relevant document
What if the results are not in a list?

- Suppose there’s only one relevant document
- Scenarios:
What if the results are not in a list?

- Suppose there’s only one relevant document
- Scenarios:
  - known-item search
What if the results are not in a list?

- Suppose there’s only one relevant document
- Scenarios:
  - known-item search
  - navigational queries
What if the results are not in a list?

- Suppose there’s only one relevant document
- Scenarios:
  - known-item search
  - navigational queries
  - looking for a fact
Mean Reciprocal Rank (MRR)

- Consider rank position, $K$, of first relevant doc
Mean Reciprocal Rank (MRR)

- Consider rank position, $K$, of first relevant doc
- Reciprocal Rank (RR) score $= \frac{1}{K}$
Mean Reciprocal Rank (MRR)

- Consider rank position, K, of first relevant doc
- Reciprocal Rank (RR) score $= \frac{1}{K}$
- MRR is the mean RR across multiple queries
Outline

Recap

Introduction

Unranked evaluation

Ranked evaluation

Benchmarks
What we need for a benchmark
What we need for a benchmark

- A collection of documents
What we need for a benchmark

- A collection of documents
  - Documents must be representative of the documents we expect to see in reality.
What we need for a benchmark

- A collection of documents
  - Documents must be representative of the documents we expect to see in reality.
- A collection of information needs
What we need for a benchmark

- A collection of documents
  - Documents must be representative of the documents we expect to see in reality.
- A collection of information needs
  - ... which we will often refer to as queries
What we need for a benchmark

- A collection of documents
  - Documents must be representative of the documents we expect to see in reality.
- A collection of information needs
  - ...which we will often refer to as queries
  - Information needs must be representative of the information needs we expect to see in reality.
What we need for a benchmark

- A collection of documents
  - Documents must be representative of the documents we expect to see in reality.
- A collection of information needs
  - … which we will often refer to as queries
  - Information needs must be representative of the information needs we expect to see in reality.
- Human relevance assessments
What we need for a benchmark

- A collection of documents
  - Documents must be representative of the documents we expect to see in reality.
- A collection of information needs
  - ...which we will often refer to as queries
  - Information needs must 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.
What we need for a benchmark

- A collection of documents
  - Documents must be representative of the documents we expect to see in reality.

- A collection of information needs
  - ...which we will often refer to as queries
  - Information needs must 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
What we need for a benchmark

- A collection of documents
  - Documents must be representative of the documents we expect to see in reality.
- A collection of information needs
  - ... which we will often refer to as queries
  - Information needs must 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 must be representative of the users we expect to see in reality.
First standard relevance benchmark: Cranfield
First standard relevance benchmark: Cranfield

- Pioneering: first testbed allowing precise quantitative measures of information retrieval effectiveness
First standard relevance benchmark: Cranfield

- Pioneering: first testbed allowing precise quantitative measures of information retrieval effectiveness
- Late 1950s, UK
First standard relevance benchmark: Cranfield

- Pioneering: first testbed allowing precise quantitative 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
First standard relevance benchmark: Cranfield

- Pioneering: first testbed allowing precise quantitative 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
Second-generation relevance benchmark: TREC

- TREC = Text Retrieval Conference (TREC)
Second-generation relevance benchmark: TREC

- TREC = Text Retrieval Conference (TREC)
- Organized by the U.S. National Institute of Standards and Technology (NIST)
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.
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 = 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
- 1.89 million documents, mainly newswire articles, 450 information needs
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
- 1.89 million documents, mainly newswire articles, 450 information needs
- No exhaustive relevance judgments – too expensive
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
- 1.89 million documents, mainly newswire articles, 450 information needs
- No exhaustive relevance judgments – too expensive
- Rather, 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.
Example of more recent benchmark: ClueWeb09
Example of more recent benchmark: ClueWeb09
Example of more recent benchmark: ClueWeb09

- 1 billion web pages
Example of more recent benchmark: ClueWeb09

- 1 billion web pages
- 25 terabytes (compressed: 5 terabyte)
Example of more recent benchmark: ClueWeb09

- 1 billion web pages
- 25 terabytes (compressed: 5 terabyte)
- Collected January/February 2009
Example of more recent benchmark: ClueWeb09

- 1 billion web pages
- 25 terabytes (compressed: 5 terabyte)
- Collected January/February 2009
- 10 languages
Example of 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)
Example of 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)
Validity of relevance assessments
Validity of relevance assessments

- Relevance assessments are only usable if they are consistent.
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.
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.
- How can we measure this consistency or agreement among judges?
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.
- How can we measure this consistency or agreement among judges?
- → Kappa measure
Kappa measure
Kappa measure

- Kappa is measure of how much judges agree or disagree.
Kappa measure

- Kappa is measure of how much judges agree or disagree.
- Corrects for chance agreement
Kappa measure

- Kappa is a measure of how much judges agree or disagree.
- Corrects for chance agreement
- $P(A) = \text{proportion of time judges agree}$
Kappa measure

- Kappa is measure of how much judges agree or disagree.
- Corrects for chance agreement
- $P(A)$ = proportion of time judges agree
- $P(E)$ = what agreement would we get by chance
Kappa measure

- Kappa is measure of how much judges agree or disagree.
- Corrects for chance agreement
- $P(A) = \text{proportion of time judges agree}$
- $P(E) = \text{what agreement would we get by chance}$
- \[ \kappa = \frac{P(A) - P(E)}{1 - P(E)} \]
Kappa measure (2)
Kappa measure (2)

- Values of $\kappa$ in the interval $[2/3, 1.0]$ are seen as acceptable.
Kappa measure (2)

- Values of $\kappa$ in the interval $[2/3, 1.0]$ are seen as acceptable.
- With smaller values: need to redesign relevance assessment methodology used etc.
Calculating the kappa statistic
Calculating the kappa statistic

<table>
<thead>
<tr>
<th></th>
<th>Judge 1</th>
<th></th>
<th>Judge 2</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>300</td>
<td>20</td>
<td>320</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>10</td>
<td>70</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
<td>90</td>
<td>400</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Calculating the kappa statistic

<table>
<thead>
<tr>
<th></th>
<th>Judge 1</th>
<th></th>
<th>Judge 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
</tr>
<tr>
<td>Yes</td>
<td>300</td>
<td>20</td>
<td>320</td>
</tr>
<tr>
<td>No</td>
<td>10</td>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
<td>90</td>
<td>400</td>
</tr>
</tbody>
</table>

- Observed proportion of the times the judges agreed

\[
P(A) = \frac{(300 + 70)}{400} = \frac{370}{400} = 0.925\]
Calculating the kappa statistic

<table>
<thead>
<tr>
<th></th>
<th>Judge 2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
</tr>
<tr>
<td>Yes</td>
<td>300</td>
<td>20</td>
<td>320</td>
</tr>
<tr>
<td>No</td>
<td>10</td>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
<td>90</td>
<td>400</td>
</tr>
</tbody>
</table>

- Observed proportion of the times the judges agreed
  \[
P(A) = \frac{300 + 70}{400} = \frac{370}{400} = 0.925
\]

- \( P(\text{nonrelevant}) = \frac{80 + 90}{400 + 400} = \frac{170}{800} = 0.2125 \)
Calculating the kappa statistic

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge 1 Yes</td>
<td>300</td>
<td>20</td>
<td>320</td>
</tr>
<tr>
<td>Judge 1 No</td>
<td>10</td>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
<td>90</td>
<td>400</td>
</tr>
</tbody>
</table>

- Observed proportion of the times the judges agreed
  \[ P(A) = \frac{300 + 70}{400} = \frac{370}{400} = 0.925 \]

- \( P(\text{nonrelevant}) = \frac{80 + 90}{(400 + 400)} = \frac{170}{800} = 0.2125 \)

- \( P(\text{relevant}) = \frac{320 + 310}{(400 + 400)} = \frac{630}{800} = 0.7878 \)
Calculating the kappa statistic

<table>
<thead>
<tr>
<th></th>
<th>Judge 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Judge 1 Yes</td>
<td>300</td>
</tr>
<tr>
<td>Judge 1 No</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
</tr>
</tbody>
</table>

- Observed proportion of the times the judges agreed
  \[ P(A) = \frac{300 + 70}{400} = \frac{370}{400} = 0.925 \]
- \[ P(\text{nonrelevant}) = \frac{80 + 90}{400 + 400} = \frac{170}{800} = 0.2125 \]
- \[ P(\text{relevant}) = \frac{320 + 310}{400 + 400} = \frac{630}{800} = 0.7878 \]
- Probability that the two judges agreed by chance
  \[ P(E) = P(\text{nonrelevant})^2 + P(\text{relevant})^2 = 0.2125^2 + 0.7878^2 = 0.665 \]
Calculating the kappa statistic

<table>
<thead>
<tr>
<th></th>
<th>Judge 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>300</td>
<td>20</td>
</tr>
<tr>
<td>No</td>
<td>10</td>
<td>70</td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
<td>90</td>
</tr>
</tbody>
</table>

- Observed proportion of the times the judges agreed
  \[ P(A) = \frac{300 + 70}{400} = \frac{370}{400} = 0.925 \]

- \[ P(\text{nonrelevant}) = \frac{80 + 90}{400 + 400} = \frac{170}{800} = 0.2125 \]

- \[ P(\text{relevant}) = \frac{320 + 310}{400 + 400} = \frac{630}{800} = 0.7878 \]

- Probability that the two judges agreed by chance
  \[ P(E) = P(\text{nonrelevant})^2 + P(\text{relevant})^2 = 0.2125^2 + 0.7878^2 = 0.665 \]

- Kappa statistic
  \[ \kappa = \frac{(P(A) - P(E))/(1 - P(E))}{1 - P(E)} = \frac{(0.925 - 0.665)/(1 - 0.665)}{1 - 0.665} = 0.776 \quad \text{(still in acceptable range)} \]
Interjudge agreement at TREC
Interjudge agreement at TREC

<table>
<thead>
<tr>
<th>Information need</th>
<th>Number of docs judged</th>
<th>Disagreements</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>211</td>
<td>6</td>
</tr>
<tr>
<td>62</td>
<td>400</td>
<td>157</td>
</tr>
<tr>
<td>67</td>
<td>400</td>
<td>68</td>
</tr>
<tr>
<td>95</td>
<td>400</td>
<td>110</td>
</tr>
<tr>
<td>127</td>
<td>400</td>
<td>106</td>
</tr>
</tbody>
</table>
Impact of interjudge disagreement
Impact of interjudge disagreement

- Judges disagree a lot. Does that mean that the results of information retrieval experiments are meaningless?
Impact of interjudge disagreement

- Judges disagree a lot. Does that mean that the results of information retrieval experiments are meaningless?
- No.
Impact of interjudge disagreement

- Judges disagree a lot. Does that mean that the results of information retrieval experiments are meaningless?
- No.
- Large impact on absolute performance numbers
Impact of interjudge disagreement

- Judges disagree a lot. Does that mean that the results of information retrieval experiments are meaningless?
- No.
- Large impact on absolute performance numbers
- Virtually no impact on ranking of systems
Impact of interjudge disagreement

- Judges disagree a lot. Does that mean that the results of information retrieval experiments are meaningless?
- No.
- Large impact on absolute performance numbers
- Virtually no impact on ranking of systems
- Suppose we want to know if algorithm A is better than algorithm B
Impact of interjudge disagreement

- Judges disagree a lot. Does that mean that the results of information retrieval experiments are meaningless?
- No.
- Large impact on absolute performance numbers
- Virtually no impact on ranking of systems
- Suppose we want to know if algorithm A is better than algorithm B
- An information retrieval experiment will give us a reliable answer to this question . . .
Impact of interjudge disagreement

- Judges disagree a lot. Does that mean that the results of information retrieval experiments are meaningless?
- No.
- Large impact on absolute performance numbers
- Virtually no impact on ranking of systems
- Suppose we want to know if algorithm A is better than algorithm B
- An information retrieval experiment will give us a reliable answer to this question . . .
- . . . even if there is a lot of disagreement between judges.
Evaluation at large search engines
Evaluation at large search engines

- Recall is difficult to measure on the web
Evaluation at large search engines

- Recall is difficult to measure on the web
- Search engines often use precision at top $k$, e.g., $k = 10$ . . .
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.
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.
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 1: clickthrough on first result
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 1: clickthrough on first result
  - 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) . . .
Evaluation at large search engines

- Recall is difficult to measure on the web
- Search engines often use precision at top $k$, e.g., $k = 10 \ldots$
- ... 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 1: clickthrough on first result
  - 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.
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 1: clickthrough on first result
  - 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.
- Example 2: A/B testing
A/B testing
A/B testing

- Purpose: Test a single innovation
A/B testing

- Purpose: Test a single innovation
- Prerequisite: You have a large search engine up and running.
A/B testing

- **Purpose**: Test a single innovation
- **Prerequisite**: You have a large search engine up and running.
- **Have most users use old system**
A/B testing

- **Purpose:** Test a single innovation
- **Prerequisite:** You have a large search engine up and running.
- **Have most users use old system**
- **Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation**
A/B testing

- Purpose: Test a single innovation
- Prerequisite: You have a large search engine up and running.
- Have most users use old system
- Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation
- Evaluate with an “automatic” measure like clickthrough on first result
A/B testing

- Purpose: Test a single innovation
- Prerequisite: You have a large search engine up and running.
- Have most users use old system
- Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation
- Evaluate with an “automatic” measure like clickthrough on first result
- Now we can directly see if the innovation does improve user happiness.
A/B testing

- Purpose: Test a single innovation
- Prerequisite: You have a large search engine up and running.
- Have most users use old system
- Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation
- Evaluate with an “automatic” measure like clickthrough on first result
- Now we can directly see if the innovation does improve user happiness.
- Probably the evaluation methodology that large search engines trust most
Critique of pure relevance
Critique of pure relevance

- We’ve defined relevance for an isolated query-document pair.
Critique of pure relevance

- We’ve defined relevance for an isolated query-document pair.
- Alternative definition: marginal relevance
Critique of pure relevance

- We’ve defined relevance for an isolated query-document pair.
- Alternative definition: marginal relevance
- The **marginal relevance** of a document at position \( k \) in the result list is the additional information it contributes over and above the information that was contained in documents \( d_1 \ldots d_{k-1} \).
Critique of pure relevance

- We’ve defined relevance for an isolated query-document pair.
- Alternative definition: marginal relevance
- The **marginal relevance** of a document at position $k$ in the result list is the additional information it contributes over and above the information that was contained in documents $d_1 \ldots d_{k-1}$.
- Exercise
Critique of pure relevance

- We’ve defined relevance for an isolated query-document pair.
- Alternative definition: marginal relevance

The **marginal relevance** of a document at position $k$ in the result list is the additional information it contributes over and above the information that was contained in documents $d_1 \ldots d_{k-1}$.

- Exercise
  - Why is marginal relevance a more realistic measure of user happiness?
Critique of pure relevance

- We’ve defined relevance for an isolated query-document pair.
- Alternative definition: marginal relevance
- The *marginal relevance* of a document at position $k$ in the result list is the additional information it contributes over and above the information that was contained in documents $d_1 \ldots d_{k-1}$.
- Exercise
  - Why is marginal relevance a more realistic measure of user happiness?
  - Give an example where a non-marginal measure like precision or recall is a misleading measure of user happiness, but marginal relevance is a good measure.
Critique of pure relevance

- We’ve defined relevance for an isolated query-document pair.
- Alternative definition: marginal relevance
- The **marginal relevance** of a document at position $k$ in the result list is the additional information it contributes over and above the information that was contained in documents $d_1 \ldots d_{k-1}$.
- **Exercise**
  - Why is marginal relevance a more realistic measure of user happiness?
  - Give an example where a non-marginal measure like precision or recall is a misleading measure of user happiness, but marginal relevance is a good measure.
  - In a practical application, what is the difficulty of using marginal measures instead of non-marginal measures?