Introduction to Information Retrieval
http://informationretrieval.org

IIR 8: Evaluation & Result Summaries

Hinrich Schütze

Center for Information and Language Processing, University of Munich

2013-05-07
Overview

1 Recap
2 Introduction
3 Unranked evaluation
4 Ranked evaluation
5 Benchmarks
6 Result summaries
Looking vs. Clicking

- Users view results one and two more often / thoroughly
- Users click most frequently on result one
Take-away today

- **Ranking** search results: why it is important (as opposed to just presenting a set of unordered Boolean results)
- **Term frequency**: This is a key ingredient for ranking.
- **Tf-idf ranking**: best known traditional ranking scheme
- **Vector space model**: Important formal model for information retrieval (along with Boolean and probabilistic models)
Why distance is a bad idea

The Euclidean distance of $\vec{q}$ and $\vec{d}_2$ is large although the distribution of terms in the query $q$ and the distribution of terms in the document $d_2$ are very similar.
Cosine similarity illustrated
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
Take-away today

- Introduction to evaluation: Measures of an IR system
- Evaluation of unranked and ranked retrieval
- Evaluation benchmarks
- Result summaries
Measures for a search engine
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
  - 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?
  - 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
  - Size of index
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
  - Size of index
  - Uncluttered UI
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
  - Size of index
  - Uncluttered UI
  - 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
  - Size of index
  - Uncluttered UI
  - Most important: **relevance**
  - (actually, maybe even more important: it’s free)
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
  - Size of index
  - Uncluttered UI
  - Most important: relevance
  - (actually, maybe even more important: it's free)
- Note that none of these is sufficient: blindingly 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
  - Size of index
  - Uncluttered UI
  - Most important: **relevance**
  - (actually, maybe even more important: it’s free)
- Note that none of these is sufficient: blindingly 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?

- 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**
- Ecommerce: 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**
- 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**
- 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
- 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
- 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.
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
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
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
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.
Relevance: query vs. information need

- User happiness can only be measured by relevance to an information need, not by relevance to queries.
- Our terminology is sloppy in these slides and in IIR: we talk about query-document relevance judgments even though we mean information-need-document relevance judgments.
Outline

1. Recap
2. Introduction
3. Unranked evaluation
4. Ranked evaluation
5. Benchmarks
6. Result summaries
Precision and recall

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

\[
\text{Precision} = \frac{\text{#(relevant items retrieved)}}{\text{#(retrieved items)}} = P(\text{relevant|retrieved})
\]
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** \((P)\) 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
You can increase recall by returning more docs.
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.

\[
F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1) PR}{\beta^2 P + R}
\]

where

\[
\beta^2 = \frac{1 - \alpha}{\alpha}
\]
A combined measure: \( F \)

- \( F \) allows us to trade off precision against recall.

\[
F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1 - \alpha}{\alpha}
\]

- \( \alpha \in [0, 1] \) and thus \( \beta^2 \in [0, \infty] \)
A combined measure: $F$

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

\[ F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \]

where \( \beta^2 = \frac{1 - \alpha}{\alpha} \)

- \( \alpha \in [0, 1] \) and thus \( \beta^2 \in [0, \infty] \)

- Most frequently used: balanced $F$ with $\beta = 1$ or $\alpha = 0.5$
A combined measure: $F$

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

\[
F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}
\]

where \( \beta^2 = \frac{1 - \alpha}{\alpha} \)

- \( \alpha \in [0, 1] \) and thus \( \beta^2 \in [0, \infty] \)
- Most frequently used: balanced $F$ with \( \beta = 1 \) or \( \alpha = 0.5 \)
  - 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.

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

where

$$\beta^2 = \frac{1 - \alpha}{\alpha}$$

- $\alpha \in [0, 1]$ and thus $\beta^2 \in [0, \infty]$.
- Most frequently used: balanced $F$ with $\beta = 1$ or $\alpha = 0.5$.
  - This is the harmonic mean of $P$ and $R$: $\frac{1}{F} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$.
- What value range of $\beta$ 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>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td>20</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>not retrieved</td>
<td>60</td>
<td>1,000,000</td>
<td>1,000,060</td>
</tr>
<tr>
<td></td>
<td>80</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>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td>20</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>not retrieved</td>
<td>60</td>
<td>1,000,000</td>
<td>1,000,060</td>
</tr>
<tr>
<td></td>
<td>80</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>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td>20</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>not retrieved</td>
<td>60</td>
<td>1,000,000</td>
<td>1,000,060</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>1,000,040</td>
<td>1,000,120</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>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td>20</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>not retrieved</td>
<td>60</td>
<td>1,000,000</td>
<td>1,000,060</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>1,000,040</td>
<td>1,000,120</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>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td>20</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>not retrieved</td>
<td>60</td>
<td>1,000,000</td>
<td>1,000,060</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>1,000,040</td>
<td>1,000,120</td>
</tr>
</tbody>
</table>

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

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

- Compute precision, recall and $F_1$ for this result set:
  - relevant    not relevant
  - retrieved   18           2
  - not retrieved 82          1,000,000,000

- 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 close to 50% for snoogle 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 close to 50% for snoogle search engine – which is too high.
- Desideratum: Punish really bad performance on either precision or recall.
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 close to 50% for snoogle 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 close to 50% for snoogle 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: 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 close to 50% for snoogle 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.
$F_1$ and other averages
$F_1$ and other averages

- We can view the harmonic mean as a kind of soft minimum
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.
Precision-recall curve
Precision-recall curve

- Precision/recall/F are measures for unranked sets.
Precision-recall curve

- Precision/recall/F are measures for **unranked sets**.
- We can easily turn set measures into measures of **ranked lists**.
Precision-recall curve

- Precision/recall/F are measures for unranked sets.
- We can easily turn set measures into measures of ranked lists.
- Just compute the set measure for each “prefix”: the top 1, top 2, top 3, top 4 etc results.
Precision-recall curve

- Precision/recall/F are measures for unranked sets.
- We can easily turn set measures into measures of ranked lists.
- Just compute the set measure for each “prefix”: the top 1, top 2, top 3, top 4 etc results
- Doing this for precision and recall gives you a precision-recall curve.
A precision-recall curve
A precision-recall curve

Each point corresponds to a result for the top $k$ ranked hits ($k = 1, 2, 3, 4, \ldots$).
A precision-recall curve

Each point corresponds to a result for the top $k$ ranked hits ($k = 1, 2, 3, 4, \ldots$).

Interpolation (in red): Take maximum of all future points.
A precision-recall curve

- Each point corresponds to a result for the top $k$ ranked hits ($k = 1, 2, 3, 4, \ldots$).
- **Interpolation (in red):** Take maximum of all future points.
- Rationale for interpolation: The user is willing to look at more stuff if both precision and recall get better.
Each point corresponds to a result for the top $k$ ranked hits ($k = 1, 2, 3, 4, \ldots$).

Interpolation (in red): Take maximum of all future points

Rationale for interpolation: The user is willing to look at more stuff if both precision and recall get better.

Questions?
11-point interpolated average precision
### 11-point interpolated average precision

<table>
<thead>
<tr>
<th>Recall</th>
<th>Interpolated Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.00</td>
</tr>
<tr>
<td>0.1</td>
<td>0.67</td>
</tr>
<tr>
<td>0.2</td>
<td>0.63</td>
</tr>
<tr>
<td>0.3</td>
<td>0.55</td>
</tr>
<tr>
<td>0.4</td>
<td>0.45</td>
</tr>
<tr>
<td>0.5</td>
<td>0.41</td>
</tr>
<tr>
<td>0.6</td>
<td>0.36</td>
</tr>
<tr>
<td>0.7</td>
<td>0.29</td>
</tr>
<tr>
<td>0.8</td>
<td>0.13</td>
</tr>
<tr>
<td>0.9</td>
<td>0.10</td>
</tr>
<tr>
<td>1.0</td>
<td>0.08</td>
</tr>
</tbody>
</table>

11-point average: \( \approx 0.425 \)
### 11-point interpolated average precision

<table>
<thead>
<tr>
<th>Recall</th>
<th>Interpolated Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.00</td>
</tr>
<tr>
<td>0.1</td>
<td>0.67</td>
</tr>
<tr>
<td>0.2</td>
<td>0.63</td>
</tr>
<tr>
<td>0.3</td>
<td>0.55</td>
</tr>
<tr>
<td>0.4</td>
<td>0.45</td>
</tr>
<tr>
<td>0.5</td>
<td>0.41</td>
</tr>
<tr>
<td>0.6</td>
<td>0.36</td>
</tr>
<tr>
<td>0.7</td>
<td>0.29</td>
</tr>
<tr>
<td>0.8</td>
<td>0.13</td>
</tr>
<tr>
<td>0.9</td>
<td>0.10</td>
</tr>
<tr>
<td>1.0</td>
<td>0.08</td>
</tr>
</tbody>
</table>

11-point average: \( \approx 0.425 \)

How can precision at 0.0 be > 0?
Averaged 11-point precision/recall graph
Averaged 11-point precision/recall graph

- Compute interpolated precision at recall levels 0.0, 0.1, 0.2, 
  …
Averaged 11-point precision/recall graph

- Compute interpolated precision at recall levels 0.0, 0.1, 0.2, ...
- Do this for each of the queries in the evaluation benchmark
Averaged 11-point precision/recall graph

- Compute interpolated precision at recall levels 0.0, 0.1, 0.2, ...
- Do this for each of the queries in the evaluation benchmark
- Average over queries
Averaged 11-point precision/recall graph

- Compute interpolated precision at recall levels 0.0, 0.1, 0.2, ...
- Do this for each of the queries in the evaluation benchmark
- Average over queries
- This measure measures performance at all recall levels.
Compute interpolated precision at recall levels 0.0, 0.1, 0.2, \ldots

Do this for each of the queries in the evaluation benchmark

Average over queries

This measure measures performance at all recall levels.

The curve is typical of performance levels at TREC.
Averaged 11-point precision/recall graph

- Compute interpolated precision at recall levels 0.0, 0.1, 0.2, ...
- Do this for each of the queries in the evaluation benchmark
- Average over queries
- This measure measures performance at all recall levels.
- The curve is typical of performance levels at TREC.
- Note that performance is not very good!
ROC curve
ROC curve

Similar to precision-recall graph
Similar to precision-recall graph

But we are only interested in the small area in the lower left corner.
Similar to precision-recall graph

But we are only interested in the small area in the lower left corner.

Precision-recall graph “blows up” this area.
Variance of measures like precision/recall
For a test collection, it is usual that a system does badly on some information needs (e.g., $P = 0.2$ at $R = 0.1$) and really well on others (e.g., $P = 0.95$ at $R = 0.1$).
Variance of measures like precision/recall

- For a test collection, it is usual that a system does badly on some information needs (e.g., $P = 0.2$ at $R = 0.1$) and really well on others (e.g., $P = 0.95$ at $R = 0.1$).
- Indeed, it is usually the case that the variance of the same system across queries is much greater than the variance of different systems on the same query.
For a test collection, it is usual that a system does badly on some information needs (e.g., $P = 0.2$ at $R = 0.1$) and really well on others (e.g., $P = 0.95$ at $R = 0.1$).

Indeed, it is usually the case that the variance of the same system across queries is much greater than the variance of different systems on the same query.

That is, there are easy information needs and hard ones.
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 should be representative of the documents we expect to see in reality.
What we need for a benchmark

- A collection of documents
  - Documents should be representative of the documents we expect to see in reality.
- A collection of information needs (often incorrectly called queries)
What we need for a benchmark

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

- A collection of information needs (often incorrectly called queries)
  - Information needs should be representative of the information needs we expect to see in reality.
What we need for a benchmark

- A collection of documents
  - Documents should be representative of the documents we expect to see in reality.
- A collection of information needs (often incorrectly called queries)
  - Information needs should be representative of the information needs we expect to see in reality.
- Human relevance assessments
What we need for a benchmark

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

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

- A collection of documents
  - Documents should be representative of the documents we expect to see in reality.
- A collection of information needs (often incorrectly called queries)
  - Information needs should be representative of the information needs we expect to see in reality.
- Human relevance assessments
  - We need to hire/pay “judges” or assessors to do this.
  - Expensive, time-consuming
  - Judges should be representative of the users we expect to see in reality.
First standard relevance benchmark: Cranfield
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 a measure of how much judges agree or disagree.
- Designed for categorical judgments
Kappa measure

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

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

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

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

\[
\kappa = \frac{P(A) - P(E)}{1 - P(E)}
\]
Kappa measure

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

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

- \( \kappa = ? \) for (i) chance agreement (ii) total agreement.
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>Judge 2 Relevance</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>Relevance 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

\[ P(A) = (300 + 70)/400 = 370/400 = 0.925 \]

Pooled marginals

\[ P(\text{nonrelevant}) = (80 + 90)/(400 + 400) = 170/800 = 0.2125 \]
\[ P(\text{relevant}) = (320 + 310)/(400 + 400) = 630/800 = 0.7878 \]

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

\[ \kappa = (P(A) - P(E))/(1 - P(E)) = (0.925 - 0.665)/(1 - 0.665) = 0.776 \text{ (still in acceptable range)} \]
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
- Supposes 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
- Supposes 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
- Supposes 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
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., $k = 10$ ...
- ... 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.
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) . . .
  - . . . but pretty reliable in the aggregate.
  - Example 2: Ongoing studies of user behavior in the lab – recall last lecture
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) . . .
  - . . . but pretty reliable in the aggregate.
  - Example 2: Ongoing studies of user behavior in the lab – recall last lecture
  - Example 3: 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 the document $d_k$ 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 the document $d_k$ 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 the document $d_k$ 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 the document $d_k$ 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 the document $d_k$ 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?
Outline

1 Recap
2 Introduction
3 Unranked evaluation
4 Ranked evaluation
5 Benchmarks
6 Result summaries
How do we present results to the user?
How do we present results to the user?

- Most often: as a list – aka “10 blue links”
How do we present results to the user?

- Most often: as a list – aka “10 blue links”
- How should each document in the list be described?
How do we present results to the user?

Most often: as a list – aka “10 blue links”

How should each document in the list be described?

This description is crucial.
How do we present results to the user?

- Most often: as a list – aka “10 blue links”
- How should each document in the list be described?
- This description is crucial.
- The user often can identify good hits (= relevant hits) based on the description.
How do we present results to the user?

- Most often: as a list – aka “10 blue links”
- How should each document in the list be described?
- This description is crucial.
- The user often can identify good hits (= relevant hits) based on the description.
- No need to actually view any document
Doc description in result list
Doc description in result list

- Most commonly: doc title, url, some metadata ...
Doc description in result list

- Most commonly: doc title, url, some metadata ...
- ...and a summary
Doc description in result list

- Most commonly: doc title, url, some metadata . . .
- . . . and a summary
- How do we “compute” the summary?
Summaries
Summaries

- Two basic kinds: (i) static (ii) dynamic
Two basic kinds: (i) static (ii) dynamic

A **static summary** of a document is always the same, regardless of the query that was issued by the user.
Summaries

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

Simplest heuristic: the first 50 or so words of the document.
Static summaries

- In typical systems, the static summary is a subset of the document.
- Simplest heuristic: the first 50 or so words of the document.
- More sophisticated: extract from each document a set of “key” sentences.
Static summaries

- In typical systems, the static summary is a subset of the document.
- Simplest heuristic: the first 50 or so words of the document.
- More sophisticated: extract from each document a set of “key” sentences.
  - Simple NLP heuristics to score each sentence.
Static summaries

- In typical systems, the static summary is a subset of the document.
- Simplest heuristic: the first 50 or so words of the document
- More sophisticated: extract from each document a set of “key” sentences
  - Simple NLP heuristics to score each sentence
  - Summary is made up of top-scoring sentences.
Static summaries

- In typical systems, the static summary is a subset of the document.
- Simplest heuristic: the first 50 or so words of the document
- More sophisticated: extract from each document a set of “key” sentences
  - Simple NLP heuristics to score each sentence
  - Summary is made up of top-scoring sentences.
  - Machine learning approach: see IIR 13
Static summaries

- In typical systems, the static summary is a subset of the document.
- Simplest heuristic: the first 50 or so words of the document.
- More sophisticated: extract from each document a set of “key” sentences
  - Simple NLP heuristics to score each sentence
  - Summary is made up of top-scoring sentences.
  - Machine learning approach: see IIR 13
- Most sophisticated: complex NLP to synthesize/generate a summary.
Static summaries

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

- Present one or more “windows” or snippets within the document that contain several of the query terms.
Dynamic summaries

- Present one or more “windows” or snippets within the document that contain several of the query terms.
- Prefer snippets in which query terms occurred as a phrase.
Dynamic summaries

- Present one or more “windows” or snippets within the document that contain several of the query terms.
- Prefer snippets in which query terms occurred as a phrase.
- Prefer snippets in which query terms occurred jointly in a small window.
Dynamic summaries

- Present one or more “windows” or snippets within the document that contain several of the query terms.
- Prefer snippets in which query terms occurred as a phrase
- Prefer snippets in which query terms occurred jointly in a small window
- The summary that is computed this way gives the entire content of the window – all terms, not just the query terms.
Google dynamic summaries for [vegetarian diet running]

No Meat Athlete | Vegetarian Running and Fitness
www.nomeatathlete.com/

Vegetarian Running and Fitness. ... (Oh, and did I mention Rich did it all on a plant-based diet?) In this episode of No Meat Athlete Radio, Doug and I had the ...
Vegetarian Recipes for Athletes - Vegetarian Shirts - How to Run Long - About

Running on a vegetarian diet – Top tips | Freedom2Train Blog
www.freedom2train.com/blog/?p=4

Nov 8, 2012 – In this article we look to tackle the issues faced by long distance runners on a vegetarian diet. By its very nature, a vegetarian diet can lead to ...

HowStuffWorks "5 Nutrition Tips for Vegetarian Runners"
www.howstuffworks.com/.../running/.../5-nutrition-tips-for-vegetarian-r...

Even without meat, you can get enough fuel to keep on running. Stockbyte/Thinkstock ...
Unfortunately, a vegetarian diet is not a panacea for runners. It could, for ...

Nutrition Guide for Vegetarian and Vegan Runners - The Running Bug
therunningbug.co.uk/.../nutrition-guide-for-vegetarian-and-vegan-runne...

Feb 28, 2012 – The Running Bug’s guide to nutrition for vegetarian and vegan ...
different types of vegetarian diet ranging from lacto-ovo-vegetarians who eat ...

Vegetarian Runner
www.vegetarianrunner.com/

Vegetarian Runner - A resource center for vegetarianism and running and how to make sure you have proper nutrition as an athlete with a vegetarian diet.
Google dynamic summaries for [vegetarian diet running]

- Good example that snippet selection is non-trivial.
Google dynamic summaries for [vegetarian diet running]

- Good example that snippet selection is non-trivial.
- Criteria: occurrence of keywords, density of keywords, coherence of snippet, number of different snippets in summary, good cutting points etc.
Generating dynamic summaries
Generating dynamic summaries

- Where do we get these other terms in the snippet from?
Generating dynamic summaries

- Where do we get these other terms in the snippet from?
- We cannot construct a dynamic summary from the positional inverted index – at least not efficiently.
Generating dynamic summaries

- Where do we get these other terms in the snippet from?
- We cannot construct a dynamic summary from the positional inverted index – at least not efficiently.
- We need to cache documents.
Generating dynamic summaries

- Where do we get these other terms in the snippet from?
- We cannot construct a dynamic summary from the positional inverted index – at least not efficiently.
- We need to cache documents.
- The positional index tells us: query term occurs at position 4378 in the document.
Generating dynamic summaries

- Where do we get these other terms in the snippet from?
- We cannot construct a dynamic summary from the positional inverted index – at least not efficiently.
- We need to cache documents.
- The positional index tells us: query term occurs at position 4378 in the document.
- Byte offset or word offset?
Generating dynamic summaries

- Where do we get these other terms in the snippet from?
- We cannot construct a dynamic summary from the positional inverted index – at least not efficiently.
- We need to cache documents.
- The positional index tells us: query term occurs at position 4378 in the document.
- Byte offset or word offset?
- Note that the cached copy can be outdated
Generating dynamic summaries

- Where do we get these other terms in the snippet from?
- We cannot construct a dynamic summary from the positional inverted index – at least not efficiently.
- We need to cache documents.
- The positional index tells us: query term occurs at position 4378 in the document.
- Byte offset or word offset?
- Note that the cached copy can be outdated
- Don’t cache very long documents – just cache a short prefix
Dynamic summaries
Dynamic summaries

- Real estate on the search result page is limited → snippets must be short . . .
Dynamic summaries

- Real estate on the search result page is limited → snippets must be short . . .
- . . .but snippets must be long enough to be meaningful.
Dynamic summaries

- Real estate on the search result page is limited → snippets must be short . . .
- . . . but snippets must be long enough to be meaningful.
- Snippets should communicate whether and how the document answers the query.
Dynamic summaries

- Real estate on the search result page is limited → snippets must be short ...
- ...but snippets must be long enough to be meaningful.
- Snippets should communicate whether and how the document answers the query.
- Ideally: linguistically well-formed snippets
Dynamic summaries

- Real estate on the search result page is limited → snippets must be short...
- ...but snippets must be long enough to be meaningful.
- Snippets should communicate whether and how the document answers the query.
- Ideally: linguistically well-formed snippets
- Ideally: the snippet should answer the query, so we don’t have to look at the document.
Dynamic summaries

- Real estate on the search result page is limited → snippets must be short . . .
- . . . but snippets must be long enough to be meaningful.
- Snippets should communicate whether and how the document answers the query.
- Ideally: linguistically well-formed snippets
- Ideally: the snippet should answer the query, so we don’t have to look at the document.
- Dynamic summaries are a big part of user happiness because . . .
Dynamic summaries

- Real estate on the search result page is limited → snippets must be short …
- … but snippets must be long enough to be meaningful.
- Snippets should communicate whether and how the document answers the query.
- Ideally: linguistically well-formed snippets
- Ideally: the snippet should answer the query, so we don’t have to look at the document.
- Dynamic summaries are a big part of user happiness because …
  - … we can quickly scan them to find the relevant document we then click on.
Dynamic summaries

- Real estate on the search result page is limited → snippets must be short . . .
- . . . but snippets must be long enough to be meaningful.
- Snippets should communicate whether and how the document answers the query.
- Ideally: linguistically well-formed snippets
- Ideally: the snippet should answer the query, so we don’t have to look at the document.
- Dynamic summaries are a big part of user happiness because . . .
  - . . . we can quickly scan them to find the relevant document we then click on.
  - . . . in many cases, we don’t have to click at all and save time.
Take-away today

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

- Chapter 8 of IIR
- Resources at http://cislmu.org
  - The TREC home page – TREC had a huge impact on information retrieval evaluation.
  - Originator of $F$-measure: Keith van Rijsbergen
  - More on A/B testing
  - Too much A/B testing at Google?
  - Tombros & Sanderson 1998: one of the first papers on dynamic summaries
  - Google VP of Engineering on search quality evaluation at Google