Diversification of Query Interpretations and Search Results

Advanced Methods of Information Retrieval
Elena Demidova
SS 2015

Materials used in the slides:
The probability ranking principle (PRP)

If an IR system’s response to each query is a ranking of documents in order of decreasing probability of relevance, the overall effectiveness of the system to its user will be maximized [1].

Let’s look at some examples and question this principle….
An example

A search engine result for the query “UPS”.

What are possible informational needs?

Which results are relevant to these informational needs?
An example

Search engine result for the query “UPS”.

Recap:

How can we quantify the quality of the search result?
Evaluation measures for ranked retrieval: precision@k

**Recap:**

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

What is the precision@5 of this result (for the informational need/s you identified)?
Evaluation measures for ranked retrieval: NDCG

NDCG: normalized discounted cumulative gain – a standard evaluation measure when graded relevance values are available.

Intuition: relevant results retrieved earlier should be rewarded.

**NDCG@k computation:**

\[
NDCG(Q, k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \frac{2^{R(j,m)} - 1}{\log_2(1+m)}
\]

**Step 1:** assess \( R(j, m) \) - the relevance of the result at position \( m \) for the query \( j \).

**Step 2:** transform the relevance of result into the gain: \( G[m] = 2^{R(j, m)} - 1 \)

**Step 3:** discount the gain of the result at position \( m \) by \( \log_2(1+m) \).

**Step 4:** cumulate the discounted gain over the top-\( k \) results (DCG)

**Step 5:** normalize the DCG using \( Z_{kj} \) - a normalization factor such that NDCG of the perfect ranking at \( k \) is 1.

**Step 6:** average over a set of queries \( Q \).

**How is NDCG@k different from the precision@k?**
Novelty and diversity in IR evaluation

nDCG@k and precision@k assume that relevance of each result can be judged in isolation, independent of other results.

Ambiguity in queries and redundancy in retrieved documents are poorly reflected by evaluation measures such as precision@k and nDCG@k.

Evaluation should systematically reward novelty and diversity.
Question answering example

85: Norwegian Cruise Lines (NCL)
  85.1: Name the ships of the NCL.
  85.2: What cruise line attempted to take over NCL in 1999?
  85.3: What is the name of the NCL’s own private island?
  85.4: How does NCL rank in size with other cruise lines?
  85.5: Why did the Grand Cayman turn away a NCL ship?
  85.6: Name so-called theme cruises promoted by NCL.

Figure 1: TREC 2005 question answering topic 85

The questions can be viewed as representatives of examples the user may be seeking.
Question answering example

85: Norwegian Cruise Lines (NCL)

85.1: Name the ships of the NCL.
85.2: What cruise line attempted to take over NCL in 1999?
85.3: What is the name of the NCL’s own private island?
85.4: How does NCL rank in size with other cruise lines?
85.5: Why did the Grand Cayman turn away a NCL ship?
85.6: Name so-called theme cruises promoted by NCL.

Figure 1: TREC 2005 question answering topic 85

<table>
<thead>
<tr>
<th>Document Title</th>
<th>85.1</th>
<th>85.2</th>
<th>85.3</th>
<th>85.4</th>
<th>85.5</th>
<th>85.6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Carnival Re-Enters Norway Bidding</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>b. NORWEIGAN CRUISE LINE SAYS OUTLOOK IS GOOD</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>c. Carnival, Star Increase NCL Stake</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>d. Carnival, Star Solidify Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>e. HOUSTON CRUISE INDUSTRY GETS BOOST WITH...</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>f. TRAVELERS WIN IN CRUISE TUG-OF-WAR</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>g. ARMCHAIR QUARTERBACKS NEED... THIS CRUISE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>h. EUROPE, CHRISTMAS ON SALE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>i. TRAVEL DEALS AND DISCOUNTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>j. HAVE IT YOUR WAY ON THIS SHIP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2: Top ten documents returned for the query “Norwegian Cruise Lines (NCL)”. The questions answered by each document are indicated.
Information nuggets

Information nugget $n$ represents any binary property of the document – e.g. an answer to a question, topicality, etc.

$N = \{n_1, \ldots, n_m\}$ is the space of all possible nuggets.

User’s information need $u$ is a set of nuggets – a subset of $N$.

The information present in a document is a set of nuggets $d$ – a subset of $N$. 
Information nuggets

Document $d$ is relevant to a query $u$ if it contains at least one relevant information nugget.
Diversity-aware evaluation measures: $\alpha$-NDCG

$\alpha$-NDCG: adjust the gain computation of NDCG to discourage redundant results

$J(d_k, i)$ – assessor judged that document $d_k$ contains nugget $n_i$.

$\alpha$ - a factor that balances relevance and novelty

$r_{i, k-1}$ – the number of documents ranked up to position $k-1$ that have been judged to contain nugget $n_i$.

Sum over all $m$ nuggets:

$$G[k] = \sum_{i=1}^{m} J(d_k, i)(1 - \alpha)^{r_{i, k-1}}$$

Other steps are equivalent to the NDCG computation.
Compute the gain vector for $\alpha$-NDCG

$J(d_k, i)$ – assessor judged that document $d_k$ contains nugget $n_i$.

$\alpha$ - a factor that balances relevance and novelty = 0.5

$r_{i, k-1}$ – the number of documents ranked up to position $k-1$ that have been judged to contain nugget $n_i$.

$i...m$ – sum over the nuggets

$$G[k] = \sum_{i=1}^{m} J(d_k, i)(1 - \alpha)^{r_{i, k-1}}$$

<table>
<thead>
<tr>
<th>Document Title</th>
<th>85.1</th>
<th>85.2</th>
<th>85.3</th>
<th>85.4</th>
<th>85.5</th>
<th>85.6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Carnival Re-Enters Norway Bidding</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>b. NORWEGIAN CRUISE LINE SAYS OUTLOOK IS GOOD</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>c. Carnival, Star Increase NCL Stake</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>d. Carnival, Star Solidify Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>e. HOUSTON CRUISE INDUSTRY GETS BOOST WITH...</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>f. TRAVELERS WIN IN CRUISE TUG-OF-WAR</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>g. ARMCHAIR QUARTERBACKS NEED... THIS CRUISE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>h. EUROPE, CHRISTMAS ON SALE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>i. TRAVEL DEALS AND DISCOUNTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>j. HAVE IT YOUR WAY ON THIS SHIP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2: Top ten documents returned for the query “Norwegian Cruise Lines (NCL)”. The questions answered by each document are indicated.
Compute the gain vector for $\alpha$-NDCG

$J(d_k, i) \ – \ \text{assessor judged that document } d_k \ \text{contains nugget } n_i.$

$\alpha \ - \ \text{a factor that balances relevance and novelty } = 0.5$

$r_{i, k-1} \ - \ \text{the number of documents ranked up to position } k-1 \ \text{that have been judged to contain nugget } n_i.$

$i\ldots m \ - \ \text{sum over the nuggets}$

\[
G[k] = \sum_{i=1}^{m} J(d_k, i)(1 - \alpha)^{r_{i, k-1}}
\]

<table>
<thead>
<tr>
<th>Document Title</th>
<th>85.1</th>
<th>85.2</th>
<th>85.3</th>
<th>85.4</th>
<th>85.5</th>
<th>85.6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Carnival Re-Enters Norway Bidding</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>b. NORWEGIAN CRUISE LINE SAYS OUTLOOK IS GOOD</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>c. Carnival, Star Increase NCL Stake</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>d. Carnival, Star Solidify Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>e. HOUSTON CRUISE INDUSTRY GETS BOOST WITH...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>f. TRAVELERS WIN IN CRUISE TUG-OF-WAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>g. ARMCHAIR QUARTERBACKS NEED... THIS CRUISE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>h. EUROPE, CHRISTMAS ON SALE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>i. TRAVEL DEALS AND DISCOUNTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>j. HAVE IT YOUR WAY ON THIS SHIP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2: Top ten documents returned for the query “Norwegian Cruise Lines (NCL)”.

The questions answered by each document are indicated.
Diversification in databases: The goals

• Materialization of results is expensive in databases
• Provide an overview of the variety of differently structured search results.
• Develop efficient algorithms and evaluation metrics for search result diversification over structured data.
• Query interpretations have clear semantics, they offer quality information for diversification
• We diversify query interpretations and avoid materializing and filtering redundant results
An example query

Keyword query:
CONSIDERATION CHRISTOPHER GUEST

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Diversification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A director CHRISTOPHER GUEST of a movie CONSIDERATION</td>
</tr>
<tr>
<td>2</td>
<td>A director CHRISTOPHER GUEST</td>
</tr>
<tr>
<td>3</td>
<td>An actor CHRISTOPHER GUEST in a movie CONSIDERATION</td>
</tr>
</tbody>
</table>

Query interpretations can be similar and return overlapping results
Recap: From keywords to structured queries

\[ K = \text{“consideration christopher guest”} \]

\[ \sigma \text{consideration} \in \text{title(Movie)} \quad \sigma \text{“christopher guest“} \in \text{name(Director)} \]

\[ T = \sigma ? \in \text{name(Director)} \bowtie \text{Directs} \bowtie \sigma ? \in \text{title(Movie)} \]

\[ Q = \sigma \text{christopher guest} \in \text{name(Director)} \bowtie \text{Directs} \bowtie \sigma \text{consideration} \in \text{title(Movie)} \]
Recap: probability of a query interpretation

\[ P(Q \mid K) = P(I,T \mid K) \]

- \( I \) is the set of keyword interpretations in \( Q \)
  
  \textit{consideration:} movie.title, “christopher guest“:director.name

- \( T \) is the template of \( Q \)
  
  \( T = \sigma_{? \in \text{name}} (\text{Director}) \bowtie \text{Directs} \bowtie \sigma_{? \in \text{title}} (\text{Movie}) \)

- We use database and query log statistics to estimate \( P(Q\mid K) \)
- We rank interpretations in order of decreasing probability of relevance
  
  \textit{Compare:} the probability ranking principle (PRP)
Relevance ranking using probability estimates

Keyword query:
CONSIDERATION CHRISTOPHER GUEST

<table>
<thead>
<tr>
<th>Rank</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A director CHRISTOPHER GUEST of a movie CONSIDERATION</td>
</tr>
<tr>
<td>2</td>
<td>A director CHRISTOPHER GUEST</td>
</tr>
<tr>
<td>3</td>
<td>An actor CHRISTOPHER GUEST in a movie CONSIDERATION</td>
</tr>
</tbody>
</table>

Although the queries are different, their results highly overlap
Diversifying query interpretations

Estimating Query Similarity

$$Sim(Q_1, Q_2) = \frac{|I_1 \cap I_2|}{|I_1 \cup I_2|}.$$ 

$I_1 = \{\text{consideration: movie.title, "christopher guest":director.name}\}$

$I_2 = \{\text{"christopher guest":director.name}\}$

$Sim(Q_1, Q_2) = 0.5$ – corresponds to the Jaccard coefficient computed on the sets of keyword interpretations

Combining Relevance and Similarity

$$Score(Q) = \lambda \cdot P(Q \mid K) - (1 - \lambda) \cdot \sum_{q \in QI} \frac{Sim(Q, q)}{|QI|}.$$ 

$\lambda$: balance relevance and novelty in the diversification procedure
Diversification algorithm

**Input:** list $L[l]$ of top-k query interpretations ranked by relevance

**Output:** list $R[r]$ of the relevant and diverse query interpretations

**Proc** Select Diverse Query Interpretations

*copy the first element of L in R;*

*while* (less than $r$ elements selected)

{ 
  //select the best candidate for $R[i]$

  *initialize* best_score=0;

  *while* (more candidates for $R[i]$in $L$){

    *if* (check score upper bound) break;

    *if* (score($L[j]$)>best_score) set new best_score;

  }

  add the candidate from $L$ with best_score to $R$

} End Proc;
Diversification

Keyword query:
CONSIDERATION CHRISTOPHER GUEST

<table>
<thead>
<tr>
<th>Rank</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A director CHRISTOPHER GUEST of a movie CONSIDERATION</td>
</tr>
<tr>
<td>2</td>
<td>An actor GUEST</td>
</tr>
<tr>
<td>3</td>
<td>As character CHRISTOPHER</td>
</tr>
</tbody>
</table>

A (better?) overview of the available results. Measures needed....
Evaluation metrics: $\alpha$-NDCG and S-Recall

$\alpha$-nDCG: a tradeoff between relevance and novelty. A document is a set of information nuggets. $\alpha$-nDCG discounts the document gain if nuggets were seen.

S-Recall is the number of unique subtopics covered by the first $k$ results, divided by the total number of subtopics.
Evaluation metrics: $\alpha$-NDCG-W and WS-Recall

An information nugget in $\alpha$-NDCG corresponds to a tuple (represented by primary key) in a database result.

In a database, relevance of pks varies a lot.

$\text{christopher guest} \in \text{name} \ (\text{Director}) \ vs. \ \text{consideration} \in \text{plottext} \ (\text{Plot})$

$\alpha$-NDCG and S-Recall do not take into account graded relevance of information nuggets or sub-topics.
Adapting gain for $\alpha$-NDCG-W

$\alpha$-NDCG-W takes into account graded relevance of subtopics

$$G[k] = \text{relevance}(Q_k) \cdot (1 - \alpha)^r$$

$r$ expresses an overlap in the results of $Q_k$ with results of the query interpretations at ranks 1…$k$-1

$$r = \sum_{p k_i \in Q_k} \sum_{j \in [1, k-1]} \left| p k_i \in Q_j \right|$$
WS-Recall

**WS-Recall** is an aggregated relevance of the subtopics covered by the top-\(k\) results divided by the maximum possible aggregated relevance when all relevant subtopics are covered:

\[
WS\text{-}\text{recall}@k = \frac{\sum_{p_k \in Q_1 \ldots k} \text{relevance}(p_k)}{\sum_{p_k \in U} \text{relevance}(p_k)}
\]

\(U\) is the set of all relevant subtopics
Evaluation

Goal: assess quality of disambiguation and diversification scheme
A user study and a set of experiments

Data sets: IMDB and Lyrics
- IMDB dataset: seven tables, 10 million records
- Lyrics dataset: five tables, 400 000 records

- Queries: Web search engine queries for movie and lyrics domains
  - 50 single- (sc) and multi-concept (mc) queries for each dataset
Accessing relevance of query interpretations

Too many interpretations. Can a query interpretation reflect the informational need represented by the query?

Ground truth collection: a user study with 16 participants

Relevance of query interpretations: accessed manually by the study participants

Relevance of pks: we transfer the relevance of a query interpretation to the pks this interpretation retrieves.
Query disambiguation: $\alpha$-nDCG-W, $\alpha=0$

In the case of $\alpha=0$, novelty of results is completely ignored, and $\alpha$-NDCG-W corresponds to the standard NDCG.

The relatively high $\alpha$-NDCG-W values for $\alpha=0$ confirm the quality of the ranking function.

Y-axis: $\alpha$-NDCG-W
X-axis: $k$ (for top-$k$ interpretations)
Rank: ranking without diversification
Div: ranking with diversification.
Diversification: $\alpha$-nDCG-W, $\alpha=0.99$

With $\alpha=0.99$, novelty becomes crucial, and results without novelty are regarded as completely redundant.

**IMDB**

Diversification performed on top of query ranking achieves significant reduction of result redundancy, while preserving retrieval quality in the majority of the cases.

**Lyrics**

Y-axis: $\alpha$-NDCG-W  
X-axis: $k$ (for top-$k$ interpretations)  
Rank: ranking without diversification  
Div: ranking with diversification.
λ: Balancing relevance and novelty

α-nDCG-W@top-5, α=0.99, Lyrics

α-NDCG-W values achieved by diversification decrease with increasing λ, until they meet α-NDCG-W of the ranking in λ=1.

The smaller the value of λ, the more visible is the impact of diversification and the more α-NDCG-W values of diversification outperform the original ranking.

With an increasing λ, relevance of query interpretations dominates over novelty and the amount of re-ranking achieved by diversification becomes smaller.

\[
Score(Q) = \lambda \cdot P(Q | K) - (1 - \lambda) \cdot \sum_{q \in QI} \frac{Sim(Q, q)}{|QI|}.
\]

Y-axis: α-NDCG-W
Rank: ranking without diversification
Div: ranking with diversification.
Summary

- The probability ranking principle (PRP)
- Evaluation measures for ranked retrieval
- Novelty and diversity in IR evaluation
- An approach for search result diversification over structured data
  - A probabilistic query disambiguation model
  - A query similarity measure
  - A greedy algorithm to obtain relevant and diverse interpretations
- Evaluation metrics to measure diversification quality
- Experiments to evaluate the diversification scheme
  - Novelty of keyword search results can be substantially improved
  - Search results obtained using the proposed diversification scheme better characterize possible answers available in the database than the results obtained by the initial relevance ranking
References and further reading

References:


Further reading:
