An Analysis of Instability in Web Search Results

Jinyoung Kim* and Vitor Carvalho**
* UMass Amherst / ** Microsoft Bing
Instability in Web Search Result

• What it is
  • Day-to-day change in Top K search results

• Why it matters
  • Difficulty in search quality evaluation
  • User perception of quality in re-finding [Teevan 09]
    • 40% of queries are re-finding
    • Re-finding is more difficult with rank changes
An Example of Instability

Daily Top 3 Results for the query ‘com’

http://www.microsoft.com/com/
http://www.cnet.com/
http://www.yahoo.com/

Range of NDCG5 : 0.22
Previous Work

- Temporal change of documents
  - 60% of documents change over a month [Adar 09]

- Temporal change of top results
  - 54% of top documents are replaced over a month [Selberg 00]
  - Change in search results are bigger than web itself

- Our work is the first large-scale analysis of the changes in top results
  - We define classes of instability
  - We introduce metrics for evaluating instability
Structural vs. Non-structural

- **Structural Changes**
  - Search engine periodically ship new ranking methods
  - Majority of search results are affected by such change

- **Non-structural Changes**
  - Search results change without such ‘events’
  - Feature values fluctuate, documents added or deleted

- We’re interested in non-structural changes
Short-term vs. Long-term

• Duration of Change
  • How many days do changes are in effect?

• Short-term Changes vs. Long-term Changes
  • Short-term : Duration of change <= 5
  • Long-term : Duration of change > 5

• We’re interested in short-term changes
Causes of Instability

- Indexing issues
  - Documents added to or removed from the index
  - Existing documents may not be found in the index

- Instability manifested in insertions/deletions

Insertion: a new document appears at top 10 results
Causes of Instability

- Ranking issues
  - Change in the ranking function itself
  - Change in the input to the ranking function

- Instability manifested in rank swaps

Swap: a pair of documents switching their relative positions
Summary of Instability Classes

- We focus on **non-structural, short-term** instabilities in Top K search results

<table>
<thead>
<tr>
<th></th>
<th>Short-term</th>
<th>Long-time</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Structural</strong></td>
<td></td>
<td>-Change in the ranking function itself</td>
</tr>
<tr>
<td><strong>Non-structural</strong></td>
<td>-Change in the ranking features</td>
<td>-Document permanently added to / deleted from the index</td>
</tr>
<tr>
<td></td>
<td>-Document temporarily not found on the index</td>
<td></td>
</tr>
</tbody>
</table>
Metrics of Instability

- **Set Overlap Ratio**
  - The ratio of set overlap between two ranked lists
  - Daily measures are averaged into a periodic measure

- **Pairwise Agreement Ratio**
  - The ratio of agreements in pairwise preferences between two ranked lists

- **Range of NDCG\(_k\)**
  \[ r\text{NDCG}_k(R_t, R_{t+n}) = \max(\text{NDCG}_k(R_t), ..., \text{NDCG}_k(R_{t+n})) \]
  \[ - \min(\text{NDCG}_k(R_t), ..., \text{NDCG}_k(R_{t+n})) \]
Instability Trends
An Analysis of Instability in Web Search Results
Data Set

• Daily snapshot of top 10 results from Bing
  • 12,600 queries, between 6/11 ~ 7/9 (4 weeks)
    • Including HRS and feature values

• Some of them were sent to Bing, Google and Y!
  • 1,000 queries, between 8/1 ~ 8/17 (2 weeks)

<table>
<thead>
<tr>
<th>Name</th>
<th>Date</th>
<th>#Queries</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bing Set</td>
<td>6/12 – 7/8</td>
<td>12,600</td>
<td>Bing internal</td>
</tr>
<tr>
<td>Big3 Set</td>
<td>8/2 – 8/16</td>
<td>1,000</td>
<td>Bing / Google / Yahoo! API</td>
</tr>
</tbody>
</table>
Classifying News Queries

- News-intent queries are treated differently
  - e.g. News documents are injected into top results
- We classified queries with news-intent
  - 924 queries out of 12,600 (7%)

- News query examples
  - lakers  blue
  - nba      cnn
  - dollar   espn
  - cardinals nhl
  - msnbc    college world series
Which Metric to Use?

- **Correlation between Metrics**
  
<table>
<thead>
<tr>
<th>Measure</th>
<th>PairAgree&lt;sub&gt;5&lt;/sub&gt;</th>
<th>vNDCG&lt;sub&gt;5&lt;/sub&gt;</th>
<th>rNDCG&lt;sub&gt;5&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overlap&lt;sub&gt;5&lt;/sub&gt;</td>
<td>0.986</td>
<td>-0.131</td>
<td>-0.347</td>
</tr>
<tr>
<td>PairAgree&lt;sub&gt;5&lt;/sub&gt;</td>
<td></td>
<td>-0.139</td>
<td>-0.362</td>
</tr>
<tr>
<td>vNDCG&lt;sub&gt;5&lt;/sub&gt;</td>
<td></td>
<td></td>
<td>0.798</td>
</tr>
</tbody>
</table>

- **We decided to use rNDCG<sub>5</sub>**
  - Correlates well with other metrics
  - Captures the fluctuation in retrieval effectiveness
  - Focuses on top positions and relevant documents
Instability Trends for Big3 Set

- Overall Instability Levels
  - Instability at lower position is greater
  - All three search engines have similar range of instability
    - Lower instability at top positions

<table>
<thead>
<tr>
<th>Measure</th>
<th>Bing</th>
<th>Google</th>
<th>Yahoo!</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overlap$_5$</td>
<td>0.958</td>
<td>0.958</td>
<td>0.948</td>
</tr>
<tr>
<td>Overlap$_{10}$</td>
<td>0.819</td>
<td>0.815</td>
<td>0.798</td>
</tr>
<tr>
<td>PairAgree$_5$</td>
<td>0.911</td>
<td>0.911</td>
<td>0.895</td>
</tr>
<tr>
<td>PairAgree$_{10}$</td>
<td>0.641</td>
<td>0.642</td>
<td>0.603</td>
</tr>
<tr>
<td>$r$NDCG$_5$</td>
<td>0.057</td>
<td>0.058</td>
<td>0.063</td>
</tr>
</tbody>
</table>
Bing vs. Yahoo vs. Google

#Queries with any change in top10 each day
Structural vs. Non-structural

#Queries with any change at top 10 each day

Period 1 (6/11~6/25)
- Major shipping event

Period 2 (6/26~7/9)
Insertions vs. Swaps

- One swap per query / one insertion for every other query
Insertions vs. Swaps

- The frequency of insertions and swaps in different rank positions
- Smaller amount of changes at top positions
Short-term vs. Long-term

- Lifespan of Insertions
  - 90% are temporary for news queries / 50% for regular queries

### News Queries (924)
- #Insertions at Day 1
- #Insertions revoked within 5 days

### Regular Queries (11,698)
Short-term vs. Long-term

- Lifespan of Swaps
  - 70% are temporary for news queries / 50% for regular queries

- #Swaps at Day1
  - News Queries (924)
  - Regular Queries (11,698)

- #Swaps revoked in 5 days
Unstable Query Examples

- Most unstable queries in Bing Set

<table>
<thead>
<tr>
<th>Query</th>
<th>rNDCG5</th>
</tr>
</thead>
<tbody>
<tr>
<td>jessica szohr vanity fair</td>
<td>0.780</td>
</tr>
<tr>
<td>mud.com</td>
<td>0.722</td>
</tr>
<tr>
<td>1980 topps lynn swann</td>
<td>0.693</td>
</tr>
<tr>
<td>yahoo dictionary</td>
<td>0.672</td>
</tr>
<tr>
<td>rbk</td>
<td>0.642</td>
</tr>
<tr>
<td>days spoilers</td>
<td>0.622</td>
</tr>
<tr>
<td>branson family fun factory</td>
<td>0.601</td>
</tr>
<tr>
<td>primo starr</td>
<td>0.592</td>
</tr>
<tr>
<td>teamliquid forums</td>
<td>0.591</td>
</tr>
</tbody>
</table>
Factors Affecting Instability
An Analysis of Instability in Web Search Results
Factors Affecting Instability

Instability = \text{Max (NDCG@5)} - \text{Min (NDCG@5)} \text{ over a 2-week period}
Factors Affecting Instability (2)

Instability = Max (NDCG@5) – Min (NDCG@5) over a 2-week period
Conclusions & Future Work
An Analysis of Instability in Web Search Results
Summary

• All major search engines suffer from instability
  • Even without structural changes

• Many changes in search results are temporary
  • 50% of insertions and swaps are revoked in 5 days

• Query characteristics affect its instability
  • Longer, tail queries suffer from more instabilities
Future Work

• More Analysis
  • To what extent instability correlates with document changes?
  • To what extent instability can affect user experience?

• Instability Prediction
  • Can we predict which queries will be most unstable?

• Instability Mitigation
  • How can we reduce the amount of instability?
Any Questions?

- Many thanks to Rich Caruana, Giridhar Kumanran, Manish Malik, Mohammad Makarechian, Jintae Kim, Fang Liu, Yeming Fang, and others!