Transparent Search Engines

Maarten de Rijke
June 15, 2017

University of Amsterdam
Talk based on joint work with Mathieu Heruer, Maartje ter Hoeve, Daan Odijk, Anne Schuth, Martijn Spitters.
Warning: Mostly work in progress
How does a search engine select which results to present?
So many criteria

- Aboutness
- Potential impact on reputation
- Importance
- Timeliness
- Quality
- Bias
- Fit with task/background
- Freshness
- Interestingness
- ...
That’s not how it works . . .

Ranker development

- Traditionally, manual labor
- Think about what it means for a document to match a query
- Combination of term frequency, document frequency, document length E.g.,

$$BM25(q, d) = \sum_{q_i \cdot tf(q_i, d)} \frac{idf(q_i) \cdot \frac{tf(q_i, d) \cdot (k_1 + 1)}{tf(q_i, d) + k_1 \cdot (1 - b + b \cdot \frac{\log_{10} d_{max}}{\text{avgdl}})}}{k_3 + qf(q, q)}$$
So many rankers . . .

- Content-based
  - Boolean model, extended Boolean model, . . .
  - Vector space model, latent semantic indexing, . . .
  - BM25 model, statistical language model, . . .
  - Span-based model, distance-aggregation model, . . .

- Structure-based
  - Document structure
  - Site structure
  - Link structure

- Based on interaction behavior
  - Number of visits, . . .
  - Clicks, . . .
So many rankers . . .

- Content-based
  - Boolean model, extended Boolean model, . . .
  - Vector space model, latent semantic indexing, . . .
  - BM25 model, statistical language model, . . .
  - Span-based model, distance-aggregation model, . . .
- Structure-based
  - Document structure
  - Site structure
  - Link structure
- Based on interaction behavior
  - Number of visits, . . .
  - Clicks, . . .
- \( \Rightarrow \) Documents represented by feature vectors
  - Features extracted for every query-document pair (e.g., score output by a traditional retrieval model)
  - Combine a large number of features
  - Incorporate new retrieval model by including the model’s output
So many learning to rank methods

- Least square retrieval function (TOIS 1989)
- Query refinement (WWW 2008)
- ListNet (ICML 2007)
- SVM-MAP (SIGIR 2007)
- Nested Ranker (SIGIR 2006)
- Pranking (NIPS 2002)
- LambdaRank (NIPS 2006)
- MPRank (ICML 2007)
- Frank (SIGIR 2007)
- MHR (SIGIR 2007)
- RankBoost (JMLR 2003)
- Learning to retrieval info (SCC 1995)
- LDM (SIGIR 2005)
- Large margin ranker (NIPS 2002)
- RankNet (ICML 2005)
- Ranking SVM (ICANN 1999)
- IRSVM (SIGIR 2006)
- Discriminative model for IR (SIGIR 2004)
- SVM Structure (JMLR 2005)
- OAP-BPM (ICML 2003)
- Subset Ranking (COLT 2006)
- GPRank (LR4IR 2007)
- QBRank (NIPS 2007)
- GBRank (SIGIR 2007)
- Constraint Ordinal Regression (ICML 2005)
- McRank (NIPS 2007)
- SoftRank (LR4IR 2007)
- AdaRank (SIGIR 2007)
- CCA (SIGIR 2007)
- ListMLE (ICML 2008)
- RankCosine (IPM 2007)
- Supervised Rank Aggregation (WWW 2007)
- Relational ranking (WWW 2008)
- Learning to order things (NIPS 1998)
- Round robin ranking (ECML 2003)
- ...
So many learning to rank methods

<table>
<thead>
<tr>
<th>Category</th>
<th>Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pointwise approach</strong></td>
<td><strong>Regression-based</strong>: Least square retrieval (TOIS 1989), Regression tree for ordinal class prediction (FI 2000), …</td>
</tr>
<tr>
<td></td>
<td><strong>Classification</strong>: Discriminative model for IR (SIGIR 2004), …</td>
</tr>
<tr>
<td></td>
<td><strong>Ordinal regression</strong>: Pranking (NIPS 2002), OAP-BPM (ECML 2003), Ranking with large margin principles (NIPS 2002), …</td>
</tr>
<tr>
<td><strong>Listwise approach</strong></td>
<td><strong>Non-measure specific</strong>: ListNet (ICML 2007), ListMLE (ICML 2008), BoltzRank (ICML 2009), …</td>
</tr>
<tr>
<td></td>
<td><strong>Measure-specific</strong>: AdaRank (SIGIR 2007), SVM-MAP (SIGIR 2007), SoftRank (LR4IR 2007), RankGP (LR4IR), …</td>
</tr>
</tbody>
</table>
Widely used ones

\[ f(\bar{x}) = \bar{w}^T \log(\bar{x}) \]
Ranker development

- Traditionally
  - Train and tune offline, then deploy online
  - Supervised learning paradigm

Why is this a good idea?

Why is this a bad idea?
Ranker development

- Traditionally
  - Train and tune offline, then deploy online
  - Supervised learning paradigm

- Move away from supervised paradigm
  - Weakly supervised rankers?
  - A search engine that improves by being used, not in a supervised manner but in a weakly supervised way?
  - Learn from the natural interactions with users
    - To evaluate rankers
    - To combine rankers
    - To create individual rankers

- Why is this a good idea?
- Why is this a bad idea?
retrieval system
agent

query
state $s_t$

implicit feedback

document list

action $a_t$

user
environment

examine
document list

evaluation
measure
reward $r_t$

generate implicit feedback
Explaining machine learned search results

In the machine learning community

- “A mismatch between the mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of interpretation.”
- “Ensemble methods like random forests pose a particular challenge, as predictions result from an aggregation or averaging procedure.”
- “Neural networks, especially with the rise of deep learning, pose perhaps the biggest challenge?what hope is there of explaining the weights learned in a multilayer neural net with a complex architecture?”
Explaining machine learned search results

In the machine learning community

• “A mismatch between the mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of interpretation.”
• “Ensemble methods like random forests pose a particular challenge, as predictions result from an aggregation or averaging procedure.”
• “Neural networks, especially with the rise of deep learning, pose perhaps the biggest challenge?what hope is there of explaining the weights learned in a multilayer neural net with a complex architecture?”

Solution directions

• Quantify the influence of input variables
• Produce understandable approximations of state-of-the-art methods
• Depending on the target audience, verbalize
Search: entity-wise

- Entity-panels show structured information about an entity
- Make related triples understandable
- Task: Given a triple from a knowledge graph, generate a sentence that describes it
• Approach
  • Given a pair of entities and a relation
  • Find and enrich candidate sentences
  • Learn to rank candidate sentences by how well they describe the relation (text, entities, relation, source)

• Where are we?
  • Significant improvements over state of the art sentence retrieval
  • Learn models per relation
  • Expanded to tail entities
  • Expanding to end-to-end learning

• Target user of explanation: non-professional end user
Other example: Classification

- Rank features used for classification by their contribution to the classifier’s decision
- Show for how many instances the features contribute to the decision
- Example: predict interest in product for ad placement; features are sites visited previously
- Target user: analyst

**Table 1. Top 10 highest ranked features according to the Evidence Counterfactual (EC), Shapley, and β.**

<table>
<thead>
<tr>
<th>Feature</th>
<th>EC</th>
<th>Shapley</th>
</tr>
</thead>
</table>

**Figure 1. Explanation curves for different ranking approaches.**

J. Moeyersoms et al. Explaining classification models built on high-dimensional sparse data. ICML 2016 Workshop on Human Interpretability in Machine Learning
Something else now.
News matters

Last year we saw a proliferation of disinformation online.

“Fake news” sites interfered with political discourse and sentiment around the world.
News matters

Last year we saw a proliferation of disinformation online. “Fake news” sites interfered with political discourse and sentiment around the world.

“Post-truth” named word of the year by Oxford Dictionaries

US election and EU referendum drive popularity of adjective describing situation ‘in which objective facts are less influential than appeals to emotion’

In the era of Donald Trump and Brexit, Oxford Dictionaries has declared “post-truth” to be its international word of the year.

Defined by the dictionary as an adjective “relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief”, editors said that use of the term “post-truth” had increased by around 26,000% in 2016 compared to last year. The spike in usage...
News matters

Last year we saw a proliferation of disinformation online. “Fake news” sites interfered with political discourse and sentiment around the world.

“Alternative facts” is a phrase used by U.S. Counselor to the President Kellyanne Conway during a Meet the Press interview on January 22, 2017, in which she defended White House Press Secretary Sean Spicer’s false statement about the attendance numbers of Donald Trump’s inauguration as President of the United States. When pressed during the interview with Chuck Todd to explain why Spicer "utter[ed] a provable falsehood", Conway stated that Spicer was giving "alternative facts". Todd responded, "Look, alternative facts are not facts. They're falsehoods."

Conway's use of the phrase "alternative facts" to describe demonstrable falsehoods was widely mocked on social media and sharply criticized by journalists and media organizations, including Dan Rather, Jill Abramson, and the Public Relations Society of America. The phrase was extensively described as Orwellian. Within four days from the interview, sales of the book 1984 had increased by 9,500%, which The New York Times and others attributed to Conway's use of the phrase, making it the number-one bestseller on Amazon.com.

Conway later defended her choice of words, defining "alternative facts" as "additional facts and alternative information".

Background

On January 21, 2017, while White House Press Secretary Sean Spicer held his first press briefing, he accused the media of deliberately underestimating the size of the crowd for President Trump's inaugural ceremony and stated that the ceremony had drawn the "largest audience to ever witness an inauguration — period — both in person and around the globe". According to available data and photographic evidence, Spicer's claims and allegations were false. Aerial images showed that the turnout for Trump's inauguration was lower than the turnout for the 2009 inauguration of Barack Obama. Spicer claimed that 420,000 people rode the DC Metro on inauguration day 2017, compared to 317,000 in 2013. It is "unclear where his 420,000 figure comes from" or what time periods he was comparing. Actual ridership figures between midnight and 11 AM were 193,000 in 2017, 317,000 in 2013. Full-day ridership was 570,557 in 2017, 792,000 in 2013.

Spicer also gave incorrect information about the use of white ground coverings during the inauguration. He stated that they were used for the first time in 2017. Actual ridership figures between midnight and 11 AM were 193,000 in 2017, 317,000 in 2013. Full-day ridership was 570,557 in 2017, 792,000 in 2013.
What can we do to help restore a common fact base?

If there is any kryptonite to false information, it’s transparency

How can technology platforms expose more information about the content people are seeing, and why they’re seeing it?

Need this visibility – it sheds light on process and origins of information and creates a structure for accountability
Explaining news search

How can we explain news search and news recommendation results?
Do people want explanations for their personalized news recommendations?

Do they care about the type of explanation?

Do explanations impact news consumer behavior?
Interacting with explanations at Blendle

(a) Single reason, visible – “Because you like reading about politics.”

(b) Single reason, invisible – “Because you like long reads and tech.”

(c) Multiple reasons, visible – “Because you read from this author before.”

(d) Multiple reasons, combined – “Because you follow De Tijd and read from these authors more often.”

(e) Bar chart – “Chosen for you based on: Author(s): Maarten Keulemans; Publication: De Volkskrant; Topic: Tech”

Figure 1: Examples reason types. Reasons are in the lines that start with “Omdat” (because), except for the bar chart. There, the reasons starts with “Voor jou gekozen”. Translations are given below each article.
Outcomes of a user study

- **Significant** majority of news users wants an explanation
- **No significant differences** in preference for one explanation type over another

Outcomes of A/B test

- **Explanations drive traffic**
- You can have your cake and eat it!
Generating explanations at Blendle

Pointwise explanations

- How does an item get its score
- I.e., which feature contributes most to the score of a log linear learning to rank model
- Treating ranking as a classification problem
Generating explanations at Blendle

Pointwise explanations

- How does an item get its score
- I.e., which feature contributes most to the score of a log linear learning to rank model
- Treating ranking as a classification problem

Listwise

- Which features contribute most to the position of a document in a ranking
- Consider all documents in the ranking and determine to which degree the ranking changes if we change the scores for a given feature
Generating explanations at Blendle

Pointwise explanations

- How does an item get its score
- I.e., which feature contributes most to the score of a log linear learning to rank model
- Treating ranking as a classification problem

Listwise

- Which features contribute most to the position of a document in a ranking
- Consider all documents in the ranking and determine to which degree the ranking changes if we change the scores for a given feature

Stay tuned . . .
Wrapping up

Explain search and recommendation results

Who are the target users

How do we assess explanations?