Ranking from Pairs and Triplets: Information Quality, Evaluation Methods and Query Complexity

Kira Radinsky
Nir Ailon
CS department, Technion, Israel
How do we know that a ranking algorithm is a good one?
- Explicit evaluation.
- Implicit evaluation.

Explicit Evaluation: We usually compare it to people explicit ranking
- Given a search query and a result
- A ranker evaluates query result relevance

What are good metrics of comparison of algorithm results?

How should we get the ranking from the people?
Rank (on a scale 0-5) how relevant this search result is to the following query:

WSDM 2011
Rank (on a scale 0-5) how relevant this search result is to the following query:

WSDM 2011
Rank (on a scale 0-5) how relevant this search result is to the following query:

WSDM 2011
Rank (on a scale 0-5) how relevant this search result is to the following query:

WSDM 2011
WSDM 2011 - Web Search and Data Mining | Google Groups
There are currently too many topics in this group that display first. To make this topic appear first, remove this option from another topic.
groups.google.com/group/wsm_nlsde/browse_thread/thread/762fbb7cb14f03d3 · Cached page

Web Search and Data Mining: The ACM WSDM Conference Series
The ACM WSDM Conference Series Web Search and Data Mining WSDM 2011. The Fourth International Conference on Web Search and Data Mining will be held in Hong Kong in February, 2011.
www.wsdm-conference.org · Cached page

cfp - WSDM2011
WSDM (pronounced "wisdom") is the premier international ACM conference covering research in the areas of search and data mining on the Web.
www.wsdm2011.org/wsdm2011/cfp · Cached page

Web Search and Data Mining: WSDM 2009 Conference
WSDM 2009. The Fourth International ACM Conference on Web Search and Data Mining will be held in Hong Kong in 2011. See the official website for WSDM 2011 for details.
www.wsdm-conference.org/2011 · Cached page
Now let’s evaluate our algorithms compared to the human ranking!
Compare using: AP/DCG/NDCG/RBP/MMR/ERR

<table>
<thead>
<tr>
<th>Algorithm Ranking</th>
<th>Human Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>
So what happened?

- **Information fed into the system:** We asked the user relevance judgments on **individual search results**

- **Evaluation:** Comparison is done on ordered lists of search results, where the **results relative position** is taken into account
Information Retrieval Detour
Listwise approach

Order the search results by relevance to the following query:

WSDM 2011
The 3 questions tackled in our work

- **Information Quality.**
  Does human response to comparative combinatorial questions on k-sets contain information that differs from that contained in relevance score responses?

- **Query Complexity.**
  Which subsets do we choose from the possible $\binom{n}{k}$ to send to raters in training? Do we need to send all $\Omega(n^k)$ possibilities?

- **Evaluation.**
  How do we evaluate an ordering of search results in testing?
Are single-relevance responses stable for ranking?

- In classic IR training design (e.g., TREC), results are evaluated by raters without any context.

  ![Query 49](image)

  ![Result1](image) ![Result1](image)

  ![Result1](image) ![Result2](image)

- **Result**: Ordering in pairs (both relevance and preference) statistically differs from singleton ordering.

- **Conclusion**: Extra context affects relevance perception.
Are triplets more stable than pairs for ranking?

- Does adding an additional context result affects the induced preference or relevance for the pair?

**Results**: The preference and relevance scores of the pairs in both setups did not differ statistically.

**Conclusion**: Ranking with pairs has similar information to ranking from triplets.
Is relevance the same as preference?

- Does relevance judgment differ from preference judgments?

Results: No statistically significant difference between rankings.

Conclusion: Relevance scores, when provided in mutual context, probably do not contradict binary preference responses.
Different context types (psychological anchoring)

- Will the relevance/preference of a result pair change in the context of an additional relevant context result as opposed to an irrelevant one?

Result: Pair-relevance and pair-preference in both control groups differ statistically, where the control with an irrelevant result give higher rankings.

Conclusion: Type of added context creates a noticeable but small difference for the purpose of comparing two result pairs.
Using obfuscated results

- Psychological experiments (Ariely) showed when 2 out of 3 alternatives are easily comparable to each other (but neither easily comparable to the third), people tend to go for the better of those 2.

- **Result:** The relevance responses did not change when users were asked to give relevance scores, but changed when asked to respond by preference.

- **Conclusion:** Relevance responses, within sufficient context, are more stable than preference responses when we care about induced preference only.
How many pairs are needed?

- You need the rater to evaluate \( O(n^2) \) pairs!
- Do we really?
Quick Reminder: Vapnik-Chervonenkis (VC) Dimension

- Let $C$ be a set of boolean functions: $f: X \rightarrow \{0,1\}$
- We say a group $S \subseteq X$ is **shattered** by $C$, if for each $R \subseteq S$, a function $g \in C$ exists, s.t.:
  - $g|_R \equiv 0$
  - $g|_{S \setminus R} \equiv 1$

\[ S = \begin{bmatrix} 0 & 1 \\ R & S \setminus R \end{bmatrix} \]

$VCdim(C)$ is the biggest group that can be shattered by $C$
Quick Reminder: Vapnik-Chervonenkis (VC) Dimension

- E.g.: \( BOX^2 = \{ box_{a,b,c,d} = [a \leq x_1 \leq b] \land [c \leq x_2 \leq d] \} \)

- \( VCDim(BOX^2) \geq 4 \). A group of size 4 exists that is shattered by the functions in \( BOX^2 \)

- \( VCDim(BOX^2) < 5 \). There is no group of size 5 that can be shattered by \( BOX^2 \).
How many pairs are needed?

- You need the rater to evaluate $O(n^2)$ pairs!
- Do we really?
- The group of all ranking permutations of size $n$:
  - $|S(V)| = n!$
- [Vapnik89]: Size of VC dimension is therefore bounded by:
  
  $\text{VC}(S(V)) \leq \log(|S(V)|) = \log(n!) = O(n \log n)$
- Or is it?
How many pairs?

- **Case 1:**
  - Assume $S'$ is a group of pairs. $|S'| = n$
  - Assume both $(u,v)$ and $(v,u)$ are in the group
  - For the assignment, that assigns both $(u,v)$ and $(v,u)$ 1, there is no hypothesis in $S(v)$ (permutation) that can achieve this.
  - $S'$ cannot be shattered.
  - VC dimension is at most $n-1$. 
How many pairs?

- **Case 2:**
  - Assume $S'$ is a group of pairs. $|S'| = n$
  - Assume both $(u,v)$ and $(v,u)$ cannot be in the group.
  - Build a graph, where the edges represent pairs in $S'$

  ![Graph Diagram]

  - As $S'$ is of size $n$, there must be a circle
  - For the assignment, that assigns the edges of the circle 1, there is no hypothesis in $S(v)$ (permutation) that is consistent with this.
  - $S'$ cannot be shattered.
  - VC dimension is at most $n-1$. To show it is exactly $n-1$, any spanning tree is equivalent to a permutation.
How many pairs?

- [Vapnik95]: For a group $A$ of $m$ pairs chosen uniformly with repetition, with probability $1 - \delta$:

$$\sup_{\pi \in S(V)} \{|L_A(V, w, \pi) - L(V, w, \pi)|\} = O\left(\sqrt{\frac{n}{m}} + \sqrt{\frac{\log 1/\delta}{m}}\right)$$

- Therefore, it is enough to make $m = O(n/\varepsilon^2)$ samples, such that the regret function $L$ is at most $\varepsilon$ (additively) worse than the optimal permutation $\Rightarrow$ Kendall III-Tau distance is also $D_{KT}(\pi, \pi^*) = O(\varepsilon)$.

- Using [Diaconis 2009] $D_{KT}$ inequality: $\Rightarrow$ We need $O(nC^2)$ pairs for the elements to be placed at distance at most $\frac{n}{C}$.  

---

Query Complexity & Evaluation
Conclusions

- There should be a match between the manner in which information is fed into the system and the manner in which it is evaluated.

- Information quality: \( \Rightarrow \) pairs are stable for ranking
  - Ordering from pairs is different than ordering from single relevance responses.
  - Ordering from pairs is not different than ordering from triplets
  - Preference and relevance judgments are better in different contexts.

- Evaluation:
  - Sampling and evaluating loss.

- Query Complexity:
  - No need to worry about the number of pairs needed (not \( O(n^2) \)).
  - We need only to sample \( nC_2 \) pairs in order to place elements at distance at most \( \frac{n}{c} \) positions on average from the optimal solution.
Future Work

- **Query Complexity (Conjecture):**
  - We conjecture that the query complexity (number of pairwise preferences we pay for) required for almost perfect optimization of the function for any practical purpose is $O(n \text{ polylog } n)$.

```plaintext
Algorithm $ALG_{\varepsilon}^{rec}(V)$
1. $n \leftarrow |V|$
2. if $n = O(1)$
   3. then return optimal solution for $V$ by exhaustive search
   4. else $\pi \leftarrow ALG_{\varepsilon}(V)$
5. $k \leftarrow$ uniformly random chosen integer in $[n/3, 2n/3]$
6. $V_L \leftarrow \{v \in V : v$ among top-$k$ in $\pi\}$
7. $V_R \leftarrow \{v \in V : v$ among bottom-$(n - k)$ in $\pi\}$
8. $\pi_L \leftarrow ALG_{\varepsilon}^{rec}(V_L)$
9. $\pi_R \leftarrow ALG_{\varepsilon}^{rec}(V_R)$
10. return concatenation of $\pi_L$ and $\pi_R$
```