Search Engine Architecture

9. Learning to Rank
Machine learning for IR ranking?

- We’ve looked at methods for ranking documents in IR
  - Cosine similarity, inverse document frequency, Pagerank, ...
- We’ve looked at methods for classifying documents using supervised machine learning classifiers
  - Logistic regression, SVMs, etc.
- Surely we can also use *machine learning* to rank the documents displayed in search results?
  - Sounds like a good idea
  - A.k.a. “machine-learned relevance” or “learning to rank”

Machine learning for IR ranking

- This “good idea” has been actively researched – and actively deployed by major web search engines – in the last 7 or so years

- Why didn’t it happen earlier?
  - Modern supervised ML has been around for about 20 years...
  - Naïve Bayes has been around for about 50 years...

Machine learning for IR ranking

• There’s some truth to the fact that the IR community wasn’t very connected to the ML community
• But there were a whole bunch of precursors:
Why weren't early attempts very successful/influential?

• Sometimes an idea just takes time to be appreciated...

• **Limited training data**
  • Especially for real world use (as opposed to writing academic papers), it was very hard to gather test collection queries and relevance judgments that are representative of real user needs and judgments on documents returned
    • This has changed, both in academia and industry

• Poor machine learning techniques

• Insufficient customization to IR problem

• Not enough features for ML to show value

Why wasn’t ML much needed?

• Traditional ranking functions in IR used a very small number of features, e.g.,
  • Term frequency
  • Inverse document frequency
  • Document length
• It was easy to tune weighting coefficients by hand
  • And people did
  • You guys did in assignment 3
    • Some of you even grid searched a bit

Why is ML needed now?

- Modern systems – especially on the Web – use a great number of features:
  - Arbitrary useful features – not a single unified model
  - Log frequency of query word in anchor text?
  - Query word in color on page?
  - # of images on page?
  - # of (out) links on page?
  - PageRank of page?
  - URL length?
  - URL contains?
  - Page edit recency?
  - Page length?
- The *New York Times* (2008-06-03) quoted Amit Singhal as saying Google was using over 200 such features.

Using Classification for Ad Hoc IR

- Collect a training corpus of \((q, d, r)\) triples
  - Relevance \(r\) is here binary (but may be multiclass, with 3–7 values)
  - Document is represented by a feature vector
    - \(\mathbf{x} = (\alpha, \omega)\)
      - \(\alpha\) is cosine similarity, \(\omega\) is minimum query window size
      - \(\omega\) is the shortest text span that includes all query words
      - Query term proximity is a **very important** new weighting factor
  - Train a machine learning model to predict the class \(r\) of a document-query pair

<table>
<thead>
<tr>
<th>example</th>
<th>docID</th>
<th>query</th>
<th>cosine score</th>
<th>(\omega)</th>
<th>judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\Phi_1)</td>
<td>37</td>
<td>linux operating system</td>
<td>0.032</td>
<td>3</td>
<td>relevant</td>
</tr>
<tr>
<td>(\Phi_2)</td>
<td>37</td>
<td>penguin logo</td>
<td>0.02</td>
<td>4</td>
<td>nonrelevant</td>
</tr>
<tr>
<td>(\Phi_3)</td>
<td>238</td>
<td>operating system</td>
<td>0.043</td>
<td>2</td>
<td>relevant</td>
</tr>
<tr>
<td>(\Phi_4)</td>
<td>238</td>
<td>runtime environment</td>
<td>0.004</td>
<td>2</td>
<td>nonrelevant</td>
</tr>
<tr>
<td>(\Phi_5)</td>
<td>1741</td>
<td>kernel layer</td>
<td>0.022</td>
<td>3</td>
<td>relevant</td>
</tr>
<tr>
<td>(\Phi_6)</td>
<td>2094</td>
<td>device driver</td>
<td>0.03</td>
<td>2</td>
<td>relevant</td>
</tr>
<tr>
<td>(\Phi_7)</td>
<td>3191</td>
<td>device driver</td>
<td>0.027</td>
<td>5</td>
<td>nonrelevant</td>
</tr>
</tbody>
</table>

Using Classification for Ad Hoc IR

• A linear score function is then
  \[ \text{Score}(d, q) = \text{Score}(\alpha, \omega) = a\alpha + b\omega + c \]
• And the linear classifier is
  Decide relevant if \( \text{Score}(d, q) > \theta \)
• ... just like text classification

Using Classification for Ad Hoc IR

Sec. 15.4.1

“Learning to Rank”

- Classification probably isn’t the right way to think about approaching ad hoc IR:
  - Classification problems: Map to a unordered set of classes
  - Regression problems: Map to a real value
  - Ordinal regression problems: Map to an ordered set of classes
    - A fairly obscure sub-branch of statistics, but what we want here
- This formulation gives extra power:
  - Relations between relevance levels are modeled
  - Documents are good versus other documents for query given collection; not an absolute scale of goodness

“Learning to Rank”

• Assume a number of categories \( C \) of relevance exist
  • These are totally ordered: \( c_1 < c_2 < \ldots < c_J \)
  • This is the ordinal regression setup

• Assume training data is available consisting of document-query pairs represented as feature vectors \( \psi_i \) and relevance ranking \( c_i \)

• We could do point-wise learning, where we try to map items of a certain relevance rank to a subinterval (e.g., Crammer et al. 2002 PRank)

• But most work does pair-wise learning, where the input is a pair of results for a query, and the class is the relevance ordering relationship between them

Point-wise Learning

- Goal is to learn a threshold to separate each rank

Pairwise Learning: The Ranking SVM

Herbrich et al. 1999, 2000; Joachims et al. 2002

- Aim is to classify instance pairs as correctly ranked or incorrectly ranked
  - This turns an ordinal regression problem back into a binary classification problem
- We want a ranking function $f$ such that
  
  $c_i > c_k \text{ iff } f(\psi_i) > f(\psi_k)$

- ... or at least one that tries to do this with minimal error
- Suppose that $f$ is a linear function
  
  $f(\psi_i) = \mathbf{w} \cdot \psi_i$

The Ranking SVM
Herbrich et al. 1999, 2000; Joachims et al. 2002

- Ranking Model: \( f(\psi_i) \)
The Ranking SVM
Herbrich et al. 1999, 2000; Joachims et al. 2002

- Then:
  \[ c_i > c_k \text{ iff } \mathbf{w} \cdot (\psi_i - \psi_k) > 0 \]
- Let us then create a new instance space from such pairs:
  \[ \Phi_u = \Phi(d_i, d_j, q) = \psi_i - \psi_k \]
  \[ z_u = +1, 0, -1 \text{ as } c_i >,=,< c_k \]
- We can build model over just cases for which \( z_u = -1 \)
- From training data \( S = \{\Phi_u\} \), we train an SVM
Demo
Modeling Problems
Cao et al. SIGIR 2006

- Correctly ordering the most relevant documents is crucial to the success of an IR system, while misordering less relevant results matters little.
  - The ranking SVM considers all ordering violations as the same.
- Some queries have many (somewhat) relevant documents, and other queries few. If we treat all pairs of results for a query equally, queries with many results will dominate the learning.
  - But actually queries with few relevant results are at least as important to do well on.

Two Problems with Ranking SVM

Cao et al. SIGIR 2006

- Cost sensitivity: negative effects of making errors on top ranked documents

  - \(d\): definitely relevant, \(p\): partially relevant, \(n\): not relevant
  
  - ranking 1: \(p \ d \ p \ n \ n \ n \ n \ n\)
  
  - ranking 2: \(d \ p \ n \ p \ n \ n \ n \ n\)

- Query normalization: number of instance pairs varies according to query

  - \(q1\): \(d \ p \ p \ n \ n \ n \ n \ n\)
  
  - \(q2\): \(d \ d \ p \ p \ p \ n \ n \ n \ n \ n\)
  
  - \(q1\) pairs: \(2^* (d, p) + 4^* (d, n) + 8^* (p, n) = 14\)
  
  - \(q2\) pairs: \(6^* (d, p) + 10^* (d, n) + 15^* (p, n) = 31\)

New Loss Function

Cao et al. SIGIR 2006

\[
\min_{\vec{w}} L(\vec{w}) = \sum_{i=1}^{l} \tau_{k(i)} \mu_{q(i)} \left[ 1 - z_i \left\langle \vec{w}, \vec{x}^{(1)}_i - \vec{x}^{(2)}_i \right\rangle \right] + \lambda \left\| \vec{w} \right\|^2
\]

- \( \tau \) weights for type of rank difference
  - Estimated empirically from effect on NDCG
- \( \mu \) weights for size of ranked result set
  - Linearly scaled versus biggest result set

Experimental Results (MSN search)

Cao et al. SIGIR 2006

Optimizing Rank-Based Measures
Yue et al. SIGIR 2007

• If we think that NDCG is a good approximation of the user’s utility function from a result ranking

• Then, let’s directly optimize this measure
  • As opposed to some proxy (weighted pairwise prefs)

• But, there are problems ...
  • Objective function no longer decomposes
    • Pairwise prefs decomposed into each pair
  • Objective function is flat or discontinuous

Discontinuity Example

- NDCG computed using rank positions
- Ranking via retrieval scores
- Slight changes to model parameters
  - Slight changes to retrieval scores
  - No change to ranking
  - No change to NDCG

\[
\begin{array}{c|ccc}
\text{Retrieval Score} & d_1 & d_2 & d_3 \\
\hline
\text{Rank} & 1 & 2 & 3 \\
\text{Relevance} & 0 & 1 & 0 \\
\end{array}
\]

NDCG = 0.63

NDCG discontinuous w.r.t model parameters!

Structural SVMs

Tsochantaridis et al. 2007

- Structural SVMs are a generalization of SVMs where the output classification space is not binary or one of a set of classes, but some complex object (such as a sequence or a parse tree)
- Here, it is a complete (weak) ranking of documents for a query
- The Structural SVM attempts to predict the complete ranking for the input query and document set
- The true labeling is a ranking where the relevant documents are all ranked in the front, e.g.,
  \[ y = [\text{green}, \text{green}, \text{red}, \text{red}, \text{red}] \]
- An incorrect labeling would be any other ranking, e.g.,
  \[ y' = [\text{green}, \text{red}, \text{green}, \text{green}, \text{red}] \]
- There are an intractable number of rankings, thus an intractable number of constraints!

Other Machine Learning Methods

• Of course:
  • Boosting: RankBoost
  • Ordinal Regression loglinear models
  • Neural Nets: RankNet
  • (Gradient-boosted) Decision Trees

Limitations of Machine Learning

- Everything that we have looked at (and most work in this area) produces linear models of features by weighting different base features.

- This contrasts with most of the clever ideas of traditional IR, which are nonlinear scalings and combinations of basic measurements:
  - log term frequency, idf, pivoted length normalization

- At present, ML is good at weighting features, but not at coming up with nonlinear scalings:
  - Designing the basic features that give good signals for ranking remains the domain of human creativity.

Summary

• The idea of learning ranking functions has been around for about 20 years
• But only recently have ML knowledge, availability of training datasets, a rich space of features, and massive computation come together to make this a hot research area
• It’s too early to give a definitive statement on what methods are best in this area ... it’s still advancing rapidly
• But machine learned ranking over many features now easily beats traditional hand-designed ranking functions in comparative evaluations
  • (in part by using the hand-designed functions as features!)
• There is every reason to think that the importance of machine learning in IR will grow in the future.