NTCIR-6 CLIR-J-J Experiments at Yahoo! Japan

16 May 2007
Sumio FUJITA
Yahoo Japan Research
Introduction

• Ranking optimization is much more complicated in commercial search engines because of many features which may affect the ranking:
  – Not only TF, DF, Document Length, ....
  – But also Title URL anchor text matching, term position, proximity, PageRank, ....
• Relevance judgment is difficult to obtain but implicit feedback is much easier to accumulate ....
• How to optimize the ranking given features and for example a set of query – document – judgment triplets.

2007/5/16
Parameter optimization by Genetic Algorithm

- Instead of learning directly ranking functions, we try to learn parameters of ranking functions.
- Optimize parameters especially sensitive to effectiveness, by using genetic algorithm.
- Simply replace human hill-climbing processes by genetic algorithm.
- In this work, we compare our official runs in NTCIR-4,5 and 6, optimized by human experts, with GA optimized experimental runs.
A Retrospective Study of Learning in Ad hoc Search Tasks

- Probabilistic Indexing by Fuhr et al., 1989.
- Logistic Regression by Gey et al., 1994.
- RankingSVM optimization using click data by Joachims, 2002.
- Genetic programming for ranking function discovery by Fan et al., 2003.
Genetic Algorithm at a glance

• Metaphor of the organic reproduction systems.
• Individuals to be examined are generated by applying genetic operations on each chromosome, representative of an individual, on which parameters to generate a particular individual are encoded.
• Each individual, representing a solution to the given task, is evaluated by a fitness function.
• By applying genetic operations on a generation of a population, set of individuals, a new generation is produced.
• Operations and evaluations are repeated until a predefined number of generations are processed.
• Adopted distributed genetic algorithm implementation by Hiroyasu et al., 2002.
Distributed Genetic Algorithm

1. Migration
2. Crossover
3. Mutation
4. Evaluation
5. Elitism
6. Selection

- Generation
- Population
- Island 0
- Island 2
- Individual

Crossover, Mutation, Evaluation, Elitism, Selection

Migration
Genetic Operations

• Migration operation:
  – moves randomly chosen individuals from an island to another island. This operation continues from the island to the next island and finally returns to the first island so that the population in each island remains the same.

• Crossover operation:
  – takes couples of individuals, chooses randomly two positions and exchanges the part between positions of each couple.

• Mutation operation:
  – consists of reversing randomly chosen one bit on each chromosome.

• Elitism:
  – Given the number of elite, the elite group of the previous generation and the same number of the best fitted individuals in the current generation are merged.

• Tournament selection operation:
  – selects randomly 4 individuals from the island and take the best fitted individual to the next generation and repeat this until the next generation is complete.
System description

- YLMS evaluation experiment system based on Lemur toolkit 2.0.1 for indexing system
- Indexing language:
  - Chasen version 2.2.9 as Japanese morphological analyzer with IPADIC dictionary version 2.5.1
- Retrieval models: TF*IDF with BM25 TF as follows

\[
    w(d,t,k1,b,k4) = (k4 + \log \frac{N}{df(t)}) \frac{(k1+1) \cdot freq(d,t)}{k1((1-b) + b \frac{dl_d}{avdl}) + freq(d,t)}
\]

\(d\) : document
\(t\) : term
\(N\) : total number of documents in the collection
\(df(t)\) : number of documents where \(t\) appears
\(freq(d,t)\) : number of occurrence of \(t\) in \(d\)
\(k1, k4, b\) : parameters
System description

• Rocchio feedback with top k documents in pilot search
  – Given k (#FB docs), #FB terms and FB pos Coeff.
• Title only runs evaluated by rigid judgment
• For NTCIR-n official runs, parameters are optimized using NTCIR-(n-1) collections by a human expert. (n=4,5,6)
• For experimental runs, we optimized by GA using the same training collections.

\[ Q' = Q + posCoeff \cdot \frac{1}{|R|} \cdot \sum_{D \in R} D \]
## Chromosome Design
- **Interface to GA** -

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Range</th>
<th>#bit</th>
</tr>
</thead>
<tbody>
<tr>
<td>K1</td>
<td>0 .. 3.0</td>
<td>8</td>
</tr>
<tr>
<td>b</td>
<td>0 .. 1.0</td>
<td>8</td>
</tr>
<tr>
<td>K4</td>
<td>0 .. 3.0</td>
<td>8</td>
</tr>
<tr>
<td>#FB Docs (Integer)</td>
<td>0 .. 31</td>
<td>5</td>
</tr>
<tr>
<td>#FB Terms (Integer)</td>
<td>0 .. 255</td>
<td>8</td>
</tr>
<tr>
<td>FB Pos Coeff</td>
<td>0 .. 2.0</td>
<td>8</td>
</tr>
</tbody>
</table>
Computational Environment

- 8 node clusters of Xeon 3.00GHz Dual CPU, 4GB RAM, PC servers
- Distributed computation by MPI
- Free BSD 4.xx operating system
- Optimization processes with 8 islands and a population of 10 for each island.
- 36 hours computation for 20 generations of optimization processes with 8 islands and a population of 10 for each island.
GA Control parameters

<table>
<thead>
<tr>
<th>Population</th>
<th>80</th>
<th>Chromosome length</th>
<th>45 or 24</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Islands</td>
<td>8</td>
<td>Mutation rate</td>
<td>1/45 or 1/24</td>
</tr>
<tr>
<td>Population / Island</td>
<td>10</td>
<td>Tournament size</td>
<td>4</td>
</tr>
<tr>
<td>Elite/Island</td>
<td>5</td>
<td>Migration rate</td>
<td>0.5</td>
</tr>
<tr>
<td>Crossover rate</td>
<td>1.0</td>
<td>Migration interval</td>
<td>5</td>
</tr>
</tbody>
</table>
Generation-Fitness curves measured by MAP

Fitness by each generation

2007/5/16
# GA Optimization

Cross validation by NTCIR-3,4,5 and 6

<table>
<thead>
<tr>
<th>Training Collections</th>
<th>Test collections</th>
<th>N3</th>
<th>N4</th>
<th>N5</th>
<th>N6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N3</td>
<td></td>
<td>0.4015</td>
<td>0.3766</td>
<td>0.3973</td>
<td>0.3022</td>
</tr>
<tr>
<td>N4</td>
<td></td>
<td>0.3532</td>
<td>0.4044</td>
<td>0.3616</td>
<td>0.3177</td>
</tr>
<tr>
<td>N5</td>
<td></td>
<td>0.3578</td>
<td>0.3800</td>
<td>0.4525</td>
<td>0.3147</td>
</tr>
<tr>
<td>N3,N4,N5</td>
<td></td>
<td>0.3941</td>
<td>0.3833</td>
<td>0.4355</td>
<td>0.3154</td>
</tr>
<tr>
<td>N6</td>
<td></td>
<td>0.3565</td>
<td>0.3892</td>
<td>0.4080</td>
<td>0.3308</td>
</tr>
</tbody>
</table>
# NTCIR-4 J-J

## Official run vs GA optimization

<table>
<thead>
<tr>
<th>NTCIR-4</th>
<th>K1</th>
<th>b</th>
<th>K4</th>
<th>#FB Docs</th>
<th>#FB Terms</th>
<th>FB pos coeff</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Official</td>
<td>1.0</td>
<td>0.35</td>
<td>1.0</td>
<td>7</td>
<td>100</td>
<td>0.1</td>
<td>0.3801</td>
</tr>
<tr>
<td>GA Opt (N3-N4)</td>
<td>0.8906</td>
<td>0.3672</td>
<td>2.1211</td>
<td>5</td>
<td>73</td>
<td>0.1797</td>
<td>0.3766</td>
</tr>
<tr>
<td>Diff</td>
<td>-0.1094</td>
<td>+0.0172</td>
<td>+1.1211</td>
<td>-2</td>
<td>-27</td>
<td>+0.0797</td>
<td>-0.0035</td>
</tr>
<tr>
<td>GA Opt (N4-N4)</td>
<td>2.625</td>
<td>0.5390</td>
<td>1.7344</td>
<td>22</td>
<td>115</td>
<td>1.1328</td>
<td>0.4044 (Fitness)</td>
</tr>
</tbody>
</table>

Official run was better!
## Official run vs GA optimization

<table>
<thead>
<tr>
<th>NTCIR-5</th>
<th>K1</th>
<th>K4</th>
<th>b</th>
<th>#FB</th>
<th>Terms</th>
<th>Docs</th>
<th>FB pos</th>
<th>coeff</th>
<th>MAP</th>
<th>#FB</th>
<th>Terms</th>
<th>Docs</th>
<th>FB pos</th>
<th>coeff</th>
<th>MAP</th>
<th>#FB</th>
<th>Terms</th>
<th>Docs</th>
<th>FB pos</th>
<th>coeff</th>
<th>MAP</th>
<th>#FB</th>
<th>Terms</th>
<th>Docs</th>
<th>FB pos</th>
<th>coeff</th>
<th>MAP</th>
<th>#FB</th>
<th>Terms</th>
<th>Docs</th>
<th>FB pos</th>
<th>coeff</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our</td>
<td>1.4</td>
<td>0.35</td>
<td>1.0</td>
<td>9</td>
<td>70</td>
<td>0.5</td>
<td>0.4193</td>
<td>0.5</td>
<td>0.3616</td>
<td>-0.577</td>
<td>0.4525</td>
<td>(Fitness)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Official</td>
<td></td>
<td>2.625</td>
<td>0.5390</td>
<td>22</td>
<td>115</td>
<td>1.1328</td>
<td>+0.6328</td>
<td>1.3828</td>
<td>10</td>
<td>1</td>
<td>1.6953</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GA Opt</td>
<td>N4-N5</td>
<td>1.225</td>
<td>0.189</td>
<td>+0.7344</td>
<td>+13</td>
<td>1.7344</td>
<td>+0.7344</td>
<td>1.3828</td>
<td>10</td>
<td>1</td>
<td>1.6953</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diff</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Official run was better!
# NTCIR-6 J-J

## Official run vs GA optimization

<table>
<thead>
<tr>
<th>NTCIR-6</th>
<th>K1</th>
<th>b</th>
<th>K4</th>
<th>#FB Docs</th>
<th>#FB Terms</th>
<th>FB pos coeff</th>
<th>MAP (Fitness)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Our Official</strong></td>
<td>1.1</td>
<td>0.4</td>
<td>1.5</td>
<td>9</td>
<td>70</td>
<td>0.8</td>
<td><strong>0.3182</strong></td>
</tr>
<tr>
<td><strong>GA Opt (N5-N6)</strong></td>
<td>1.1484</td>
<td>0.3945</td>
<td>1.3828</td>
<td>10</td>
<td>104</td>
<td>1.6953</td>
<td><strong>0.3147</strong></td>
</tr>
<tr>
<td><strong>Diff</strong></td>
<td>+0.0484</td>
<td>-0.0055</td>
<td>-0.1172</td>
<td>+1</td>
<td>+34</td>
<td>+0.8953</td>
<td><strong>-0.0035</strong></td>
</tr>
<tr>
<td><strong>GA Opt (N6-N6)</strong></td>
<td>1.3594</td>
<td>0.5508</td>
<td>2.4492</td>
<td>12</td>
<td>71</td>
<td>0.6953</td>
<td><strong>0.3308</strong> (Fitness)</td>
</tr>
</tbody>
</table>

Official run was better!
Observations

• GA achieves good fitness against training collections.
  – Better than human optimized best official runs
  – N4-N4-GA(0.4044) > N4-Official(0.3801) : +6.4% (No data)
  – N5-N5-GA(0.4525) > N5-Official(0.4193) : +7.9% (Sig. p=0.05)
  – N6-N6-GA(0.3308) > N6-Official(0.3182) : +4.0% (Not sig.)

• But not as good as human experts against test collections.
  – N4-Official(0.3801) > N3-N4-GA(0.3766) : +0.9% (No data)
  – N5-Official(0.4193) > N4-N5-GA(0.3616) : +16.0% (Sig. p=0.05)
  – N6-Official(0.3182) > N5-N6-GA(0.3147) : +1.1% (Not sig.)

• Large difference between fitness and test run evaluation
  – N4-N4-GA(0.4044) > N3-N4-GA(0.3766) : +7.4% (Sig. p=0.05)
  – N5-N5-GA(0.4525) > N4-N5-GA(0.3616) : +25.1% (Sig. p=0.05)
  – N6-N6-GA(0.3308) > N5-N6-GA(0.3147) : +5.1% (Sig. p=0.05)
GA Optimization without feedback
Cross validation by NTCIR-3,4,5 and 6

<table>
<thead>
<tr>
<th>Training Collections</th>
<th>Test collections</th>
<th>N3</th>
<th>N4</th>
<th>N5</th>
<th>N6</th>
</tr>
</thead>
<tbody>
<tr>
<td>N3</td>
<td>0.3413</td>
<td>0.3190</td>
<td>0.3216</td>
<td>0.2520</td>
<td></td>
</tr>
<tr>
<td>N4</td>
<td>0.3368</td>
<td>0.3215</td>
<td>0.3237</td>
<td>0.2495</td>
<td></td>
</tr>
<tr>
<td>N5</td>
<td>0.3326</td>
<td>0.3148</td>
<td>0.3300</td>
<td>0.2432</td>
<td></td>
</tr>
<tr>
<td>N3,N4,N5</td>
<td>0.3407</td>
<td>0.3187</td>
<td>0.3267</td>
<td>0.2484</td>
<td></td>
</tr>
<tr>
<td>N6</td>
<td>0.3379</td>
<td>0.3194</td>
<td>0.3217</td>
<td>0.2539</td>
<td></td>
</tr>
</tbody>
</table>
Observations

• Differences between fitness and test evaluation are much smaller.
  – N4-N4-GA-NoFB(0.3215) > N3-N4-GA-NoFB (0.3190) : +0.8% (Not sig.)
  – N5-N5-GA-NoFB(0.3300) > N4-N5-GA-NoFB (0.3237) : +1.9% (Not Sig.)
  – N6-N6-GA-NoFB(0.2539) > N5-N6-GA-NoFB (0.2432) : +4.4% (Sig. p=0.05)

• Without feedback, GA optimized runs may be as good as human experts.

• As a human expert, I sometimes worked more than 36 hours for optimizing official runs!
Conclusions

• Automatic optimization by GA achieves good fitness but in the NTCIR experimental contexts, it is not as good as human experts.
• Without feedback, GA optimized runs may be as good as human experts.
• Feedback parameters largely affect performance but learning feedback parameters causes overfitting.
• Without feedback, GA optimized parameters perform as effective to test collections as training collections.
• In the operating commercial search engines, GA optimization is used for some parameters, that are insensitive to overfitting.
• In commercial search engine optimization, with more training examples and more parameters, GA optimization is probably more effective than in NTCIR contexts.