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.

1. Council on Library and Information ... - このページを和訳
   CLIR identifies issues that affect the welfare and prospects of libraries and archives; convenes individuals and ...
   www.**clir**.org - キャッシュ
2. **CLIR** Publications - このページを和訳
   CLIR Issues. Annual Report. Other Resources ... **CLIR** publishes newsletters, reports, and other occasional items. ...
   www.**clir**.org/pubs/pubs.html - キャッシュ
3. Cross-language Information Retrieval ... (PDF)
   ... for Japanese-Chinese IR and **CLIR**, which takes ... (**CLIR**) is a special case of information retrieval. **CLIR** ...
   cl.aist-nara.ac.jp/thesis/dthesis-hasan00.pdf - htmlで見る
4. Clair ~フリー素材クレール
   透明感のある明るい素材。ガラス調、女の子向けの淡い色のホームページ素材を配布しています。
   clair.jpnet.biz - キャッシュ
5. NTCIR Workshop 4 Meeting Evaluation Form
   a. Good (**CLIR**, CLQA, PATENT, QAC, WEB, MuST) b. Bad (**CLIR**, ...
   research.nii.ac.jp/ntcir-ws5/Ev_form.html - キャッシュ
6. digital poster register number
   CLIR-1. DAEDALUS - Data Decisions and ... **CLIR**-12. Queensland University of ... **CLIR**-15. The Hong Kong Polytechnic ...
   research.nii.ac.jp/ntcir-ws5/sanka-id-no.html - キャッシュ
7. kimura, yusuke
   Such a retrieval is called cross-language information retrieval (**CLIR**) ... as I know, there is no **CLIR** for images in WWW. ...
   www.ics.es.osaka-u.ac.jp/paper/theses/2004/BC/Kimura,Yusuke - キャッシュ

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.

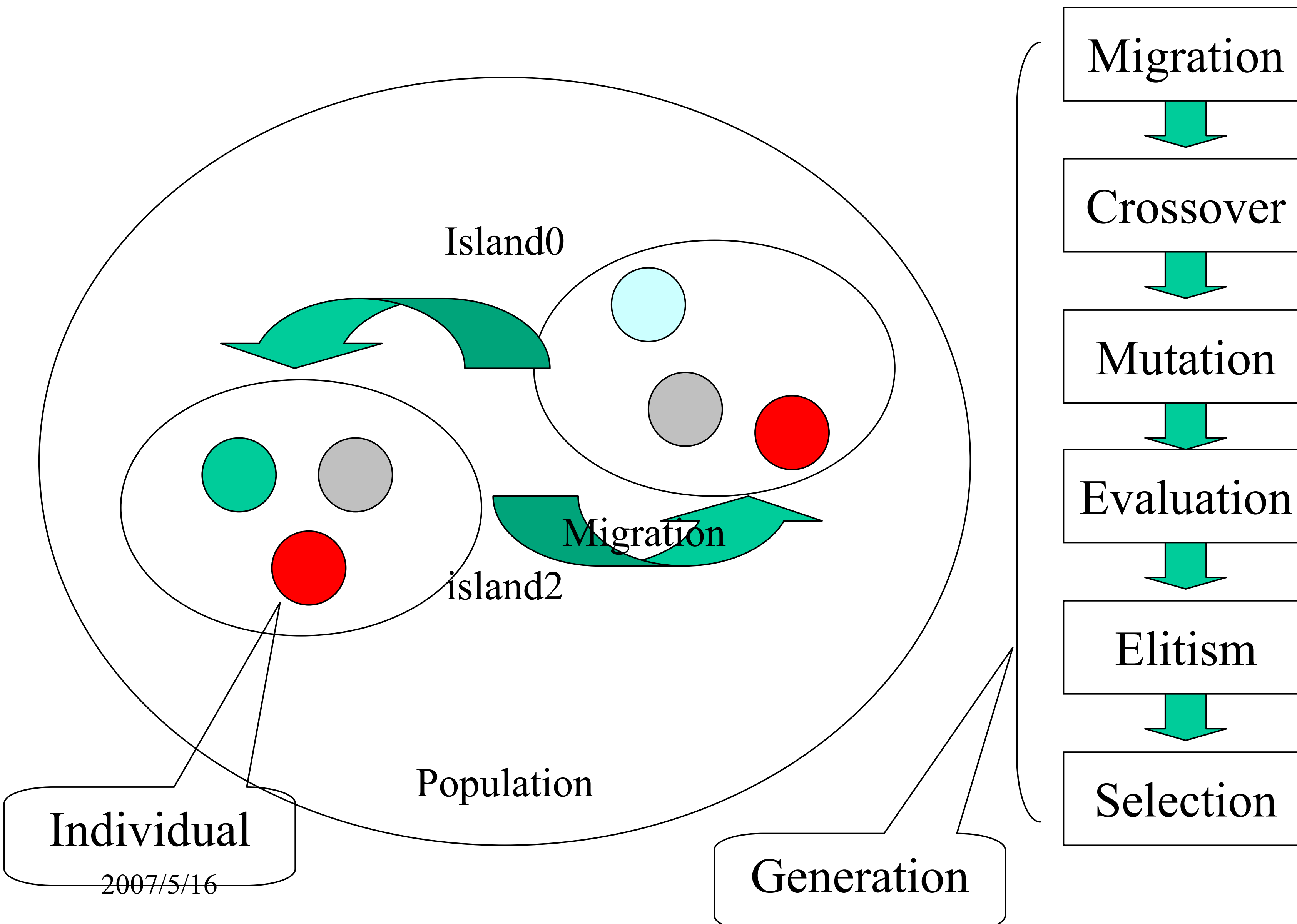
A Retrospective Study of Learning in Ad hoc Search Tasks

- Probabilistic Ranking Principle by Robertson, 1977.
- Genetic Optimization of Document Indexing by Gordon, 1988.
- 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



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

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

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-

Parameter	Range	#bit
K1	0 .. 3.0	8
b	0 .. 1.0	8
K4	0 .. 3.0	8
#FB Docs (Integer)	0 .. 31	5
#FB Terms (Integer)	0 .. 255	8
FB Pos Coeff	0 .. 2.0	8

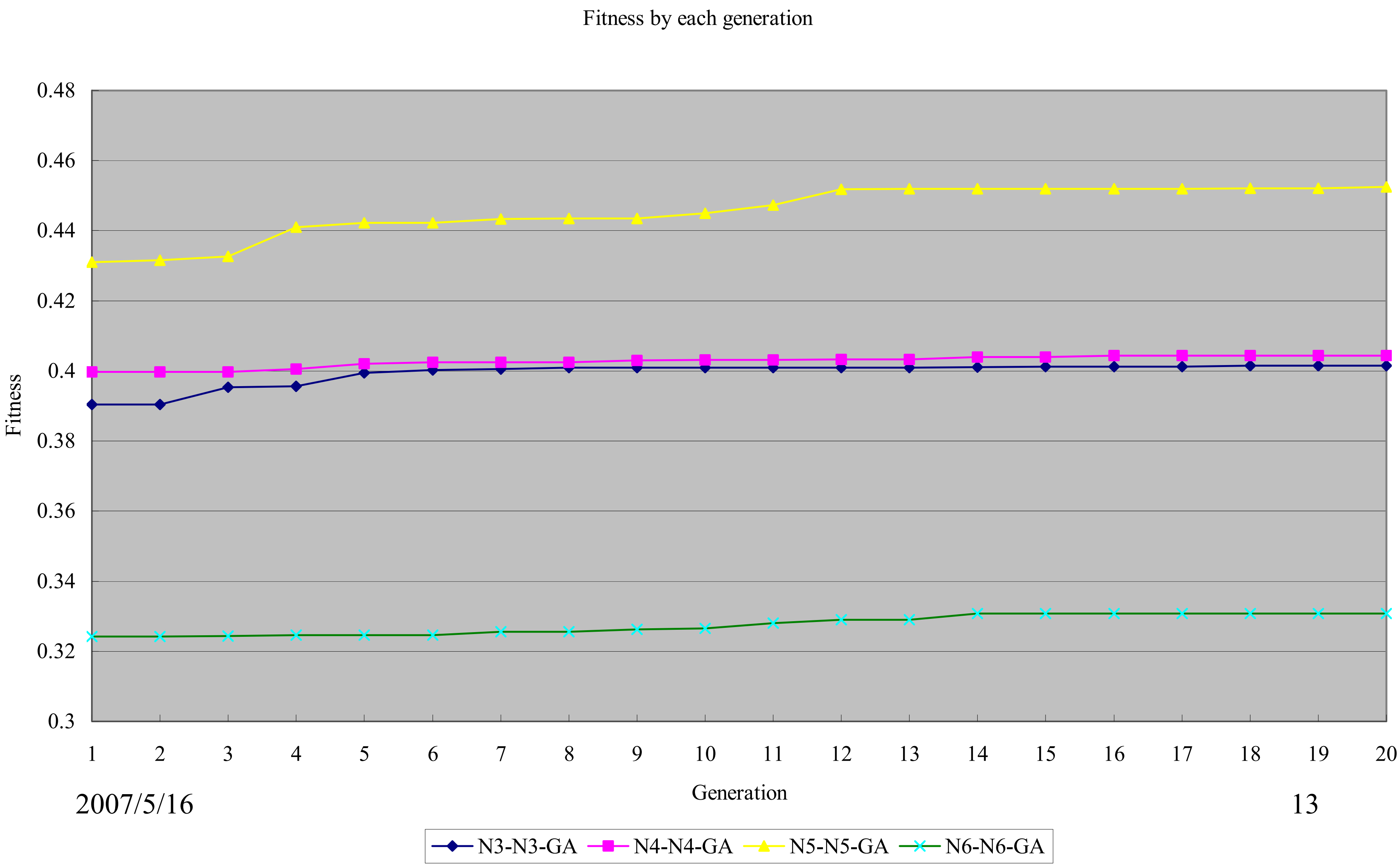
Computational Environment

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

GA Control parameters

Population	**80**	**Chromosome length**	**45 or 24**
Number of Islands	**8**	**Mutation rate**	**1/45 or 1/24**
Population / Island	**10**	**Tournament size**	**4**
Elite/Island	**5**	**Migration rate**	**0.5**
Crossover rate	**1.0**	**Migration interval**	**5**

Generation-Fitness curves measured by MAP



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

		Test collections			
		N3	N4	N5	N6
Training Collections	N3	0.4015	0.3766	0.3973	0.3022
	N4	0.3532	0.4044	0.3616	0.3177
	N5	0.3578	0.3800	0.4525	0.3147
	N3,N4,N5	0.3941	0.3833	0.4355	0.3154
	N6	0.3565	0.3892	0.4080	0.3308

NTCIR-4 J-J
Official run vs GA optimization

Official run was better!

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

NTCIR-5 J-J
Official run vs GA optimization

Official run was better!

NTCIR-5	K1	b	K4	#FB Docs	#FB Terms	FB pos coeff	MAP
Our Official	1.4	0.35	1.0	9	70	0.5	0.4193
GA Opt (N4-N5)	2.625	0.5390	1.7344	22	115	1.1328	0.3616
Diff	+1.225	+0.189	+0.7344	+13	+45	+0.6328	-0.0577
GA Opt (N5-N5)	1.1484	0.3945	1.3828	10	104	1.6953	0.4525 (Fitness)

Official run was better!

NTCIR-6 J-J
Official run vs GA optimization

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

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

		Test collections			
		N3	N4	N5	N6
Training Collections	N3	0.3413	0.3190	0.3216	0.2520
	N4	0.3368	0.3215	0.3237	0.2495
	N5	0.3326	0.3148	0.3300	0.2432
	N3,N4,N5	0.3407	0.3187	0.3267	0.2484
	N6	0.3379	0.3194	0.3217	0.2539

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.