YJRS at the NTCIR-13 OpenLiveQ Task

December 8, 2017

Tomohiro Manabe, Akiomi Nishida, Sumio Fujita
Yahoo Japan Corporation

YAHOO! JAPAN

Abstract

We started from the baseline method.
Our modifications are:

- Addition of BM25F features
  - extended for numeric field values as well as term frequencies,
- five-fold cross validation,
- and nDCG@10 as the objective function.

Our method performed well in offline and online tests due to its robustness.

YAHOO! JAPAN

Our Approaches

- Baseline method
- BM25F as ranking feature
- Extended BM25F
- Cross validation
- Objective function

YAHOO! JAPAN

Baseline Method

Baseline Method

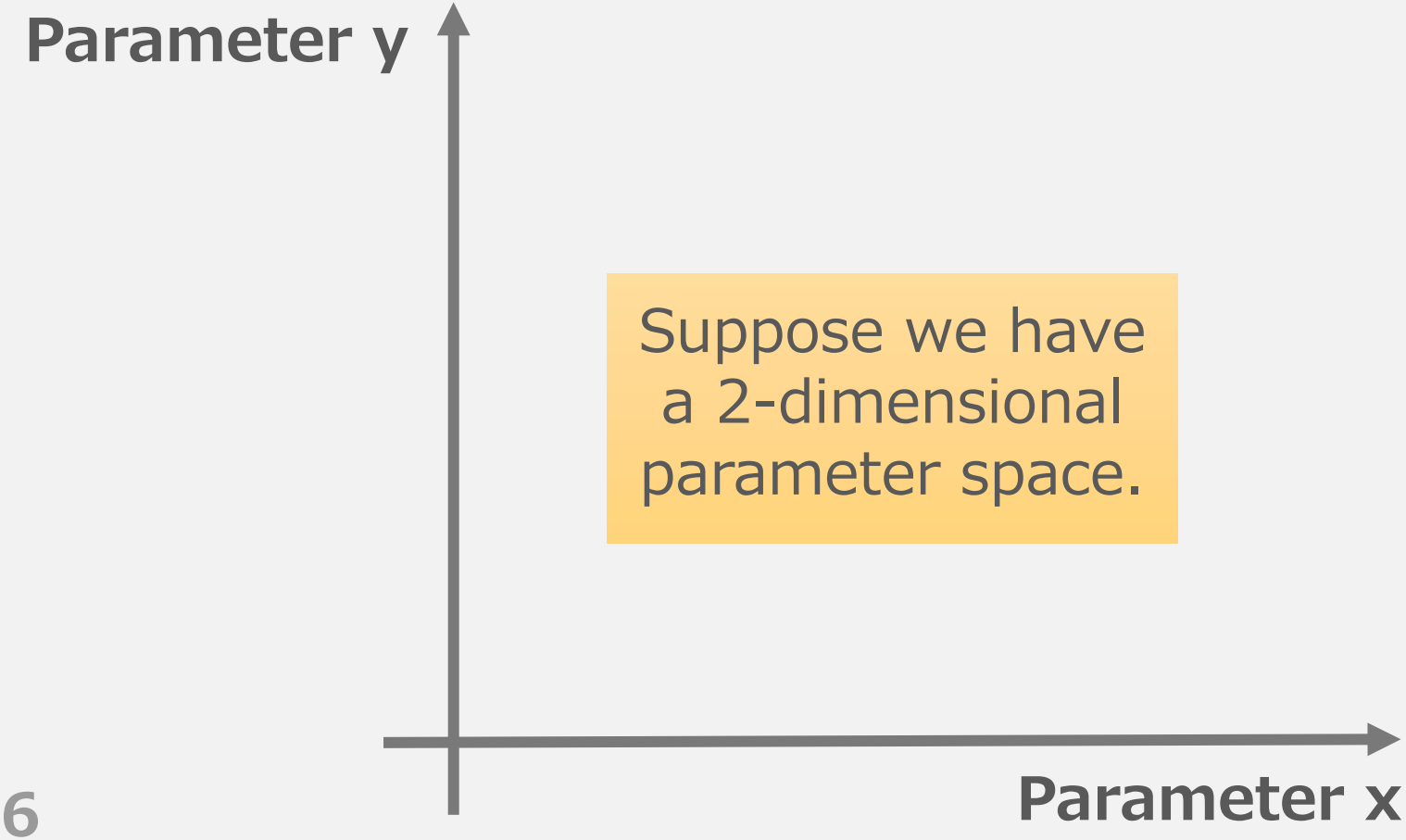
is a linear combination of 77 features:

- 68 (= 4 fields x 17) textual features (variations of TFIDF, LM, BM25),
- and 9 numeric or binary features (# of answers or PVs, baseline rank, date, open to answer or not, ……).

The weights are optimized by the Coordinate Ascent method (CA).

Baseline Method (cont'd)

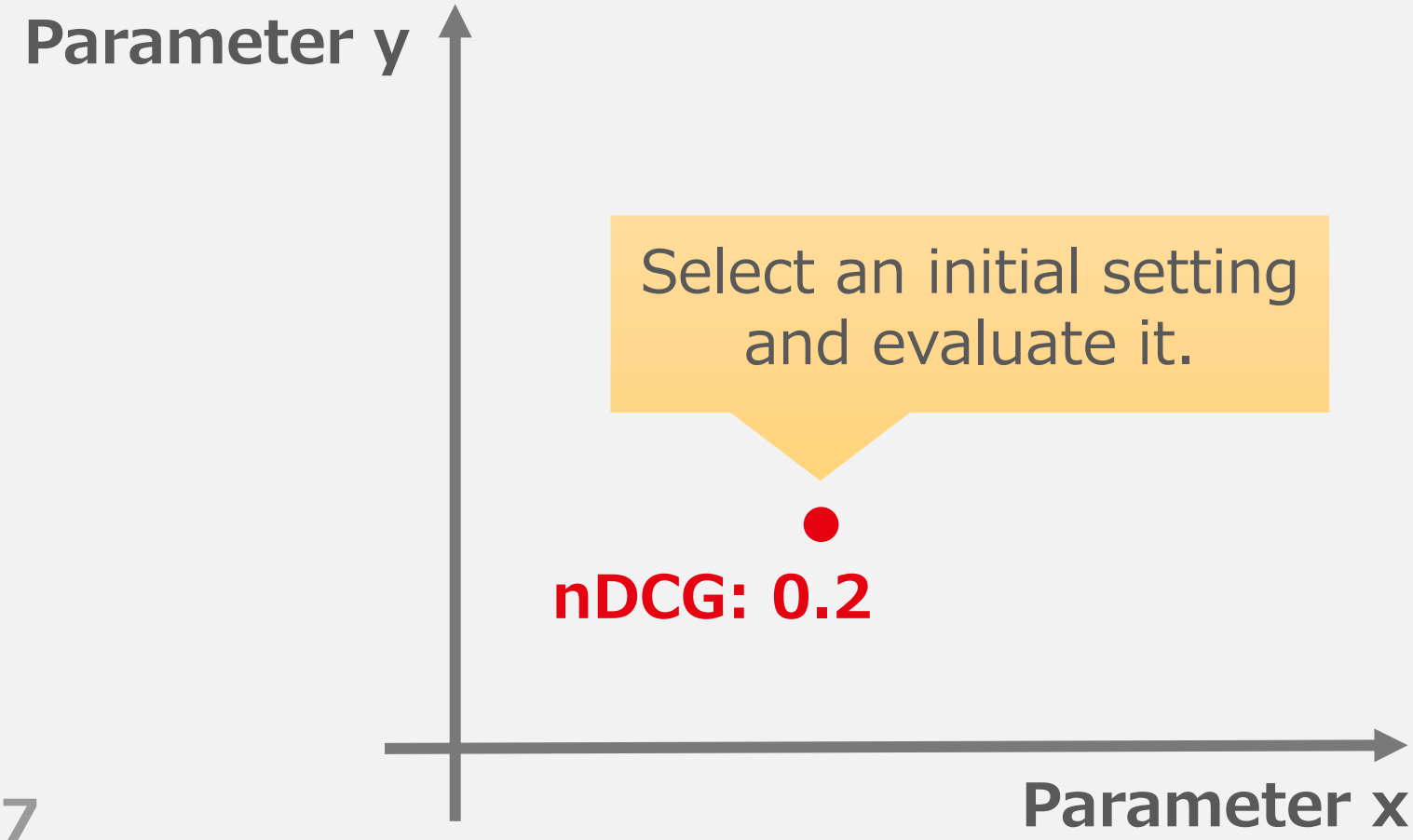
CA optimizes one parameter at once.



Suppose we have a 2-dimensional parameter space.

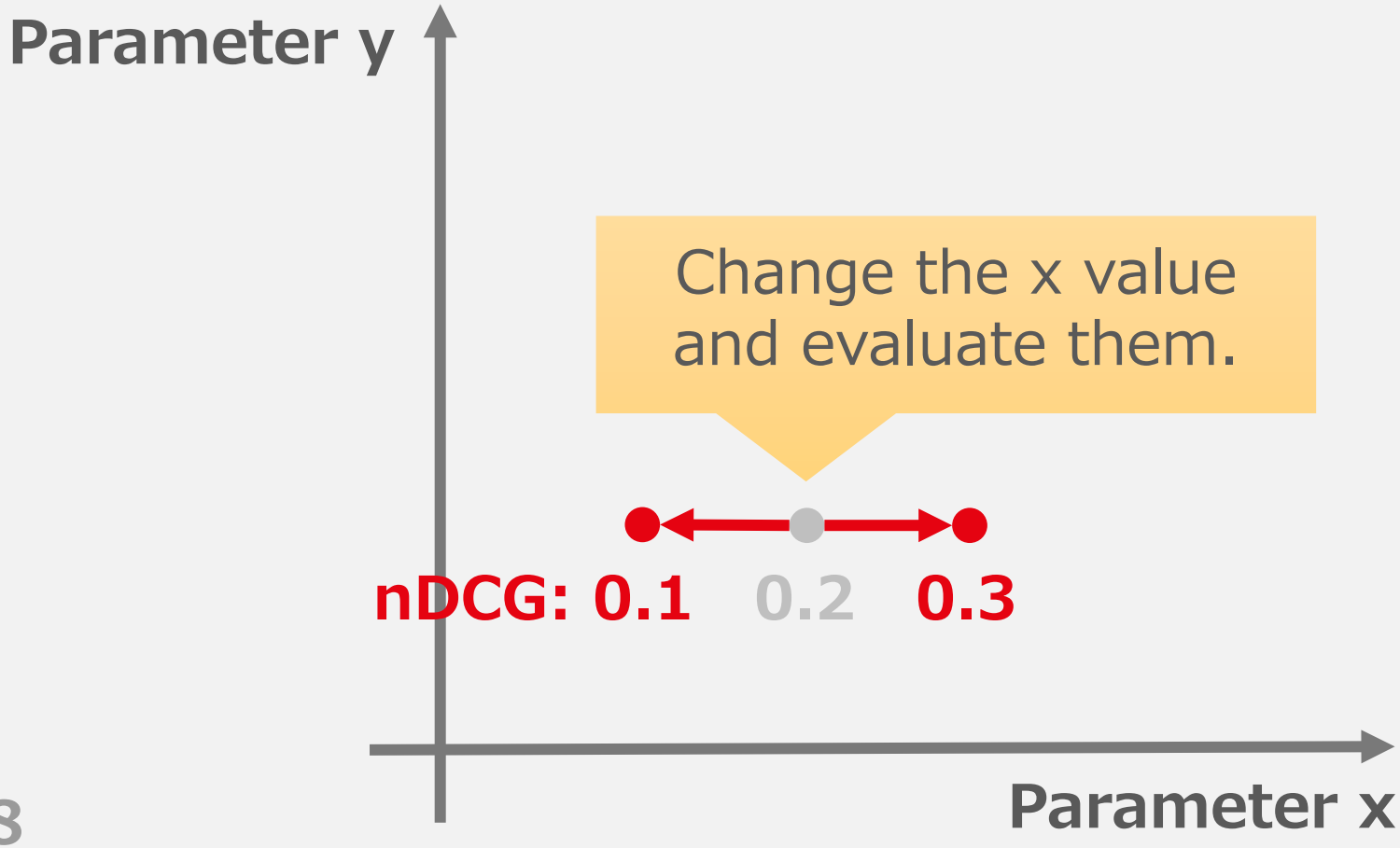
Baseline Method (cont'd)

CA optimizes one parameter at once.



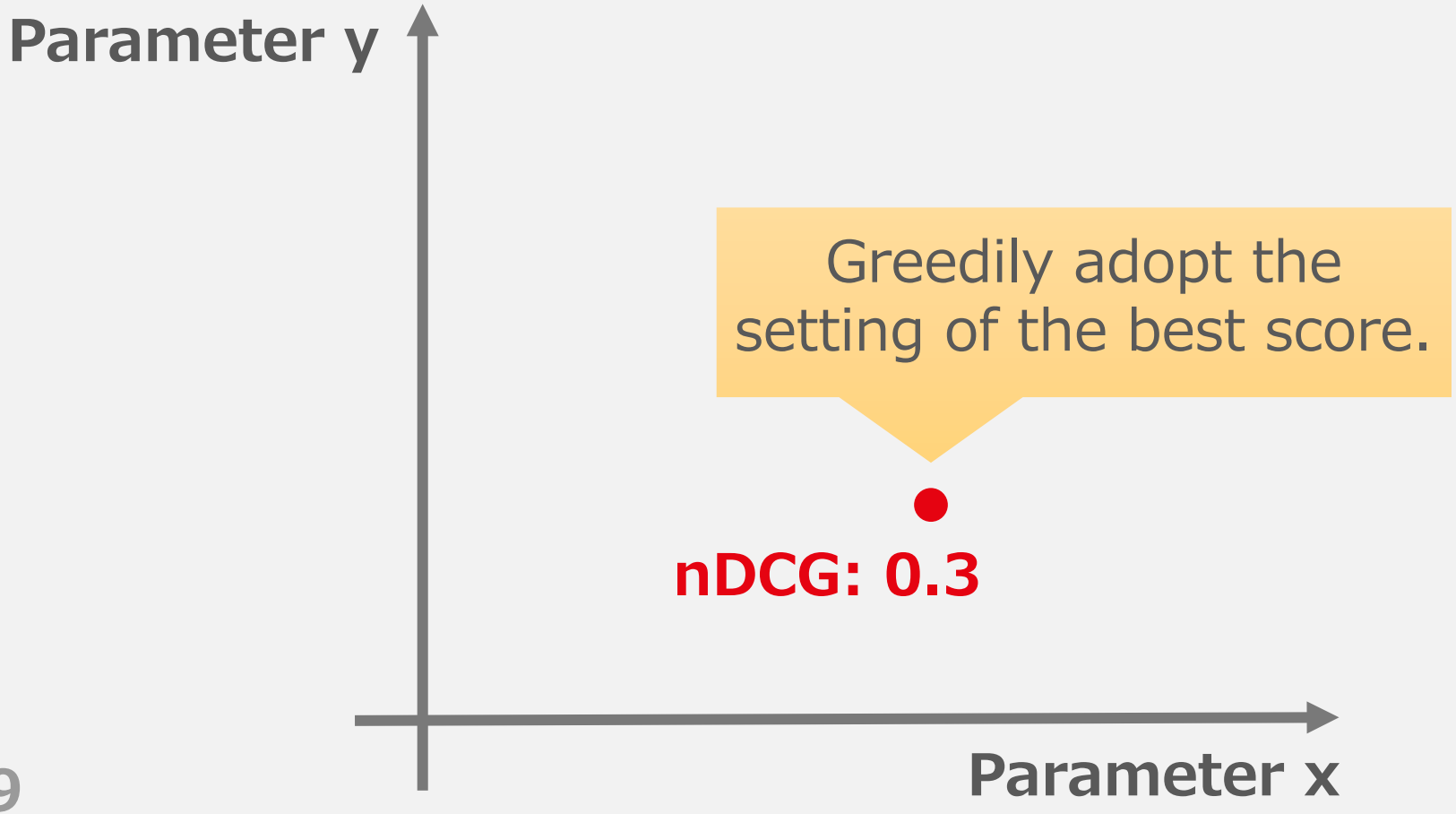
Baseline Method (cont'd)

CA optimizes one parameter at once.



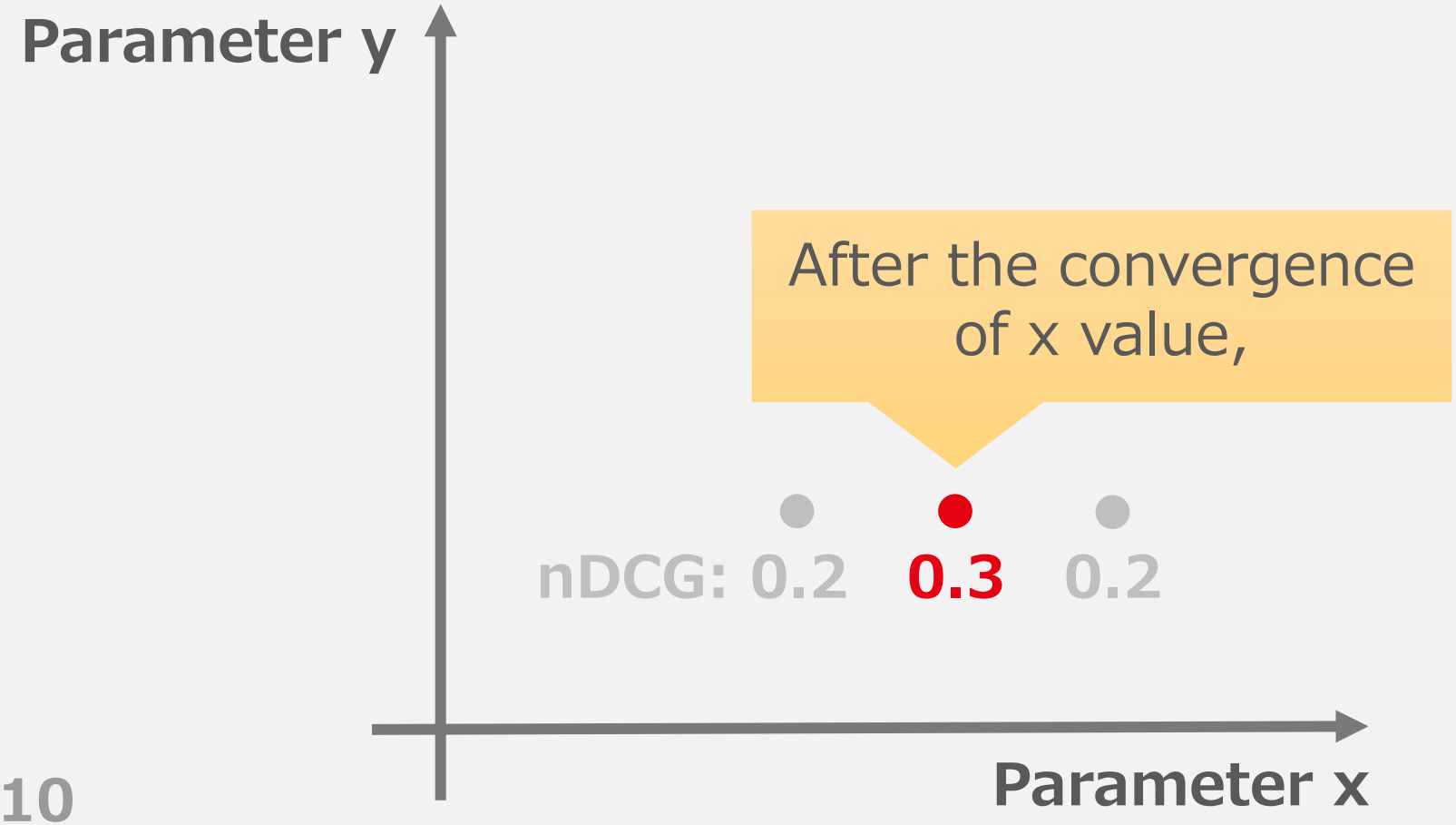
Baseline Method (cont'd)

CA optimizes one parameter at once.



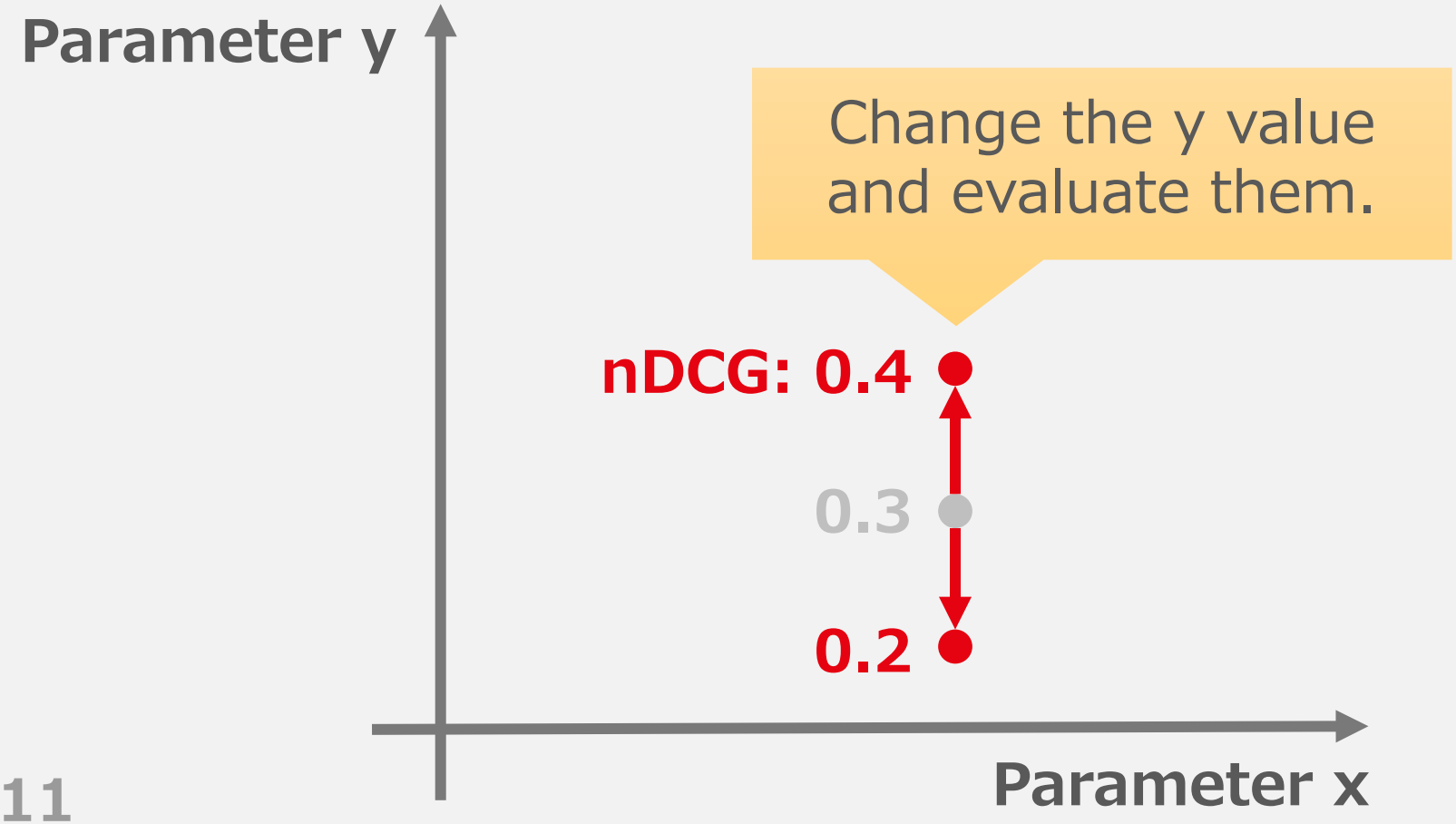
Baseline Method (cont'd)

CA optimizes one parameter at once.



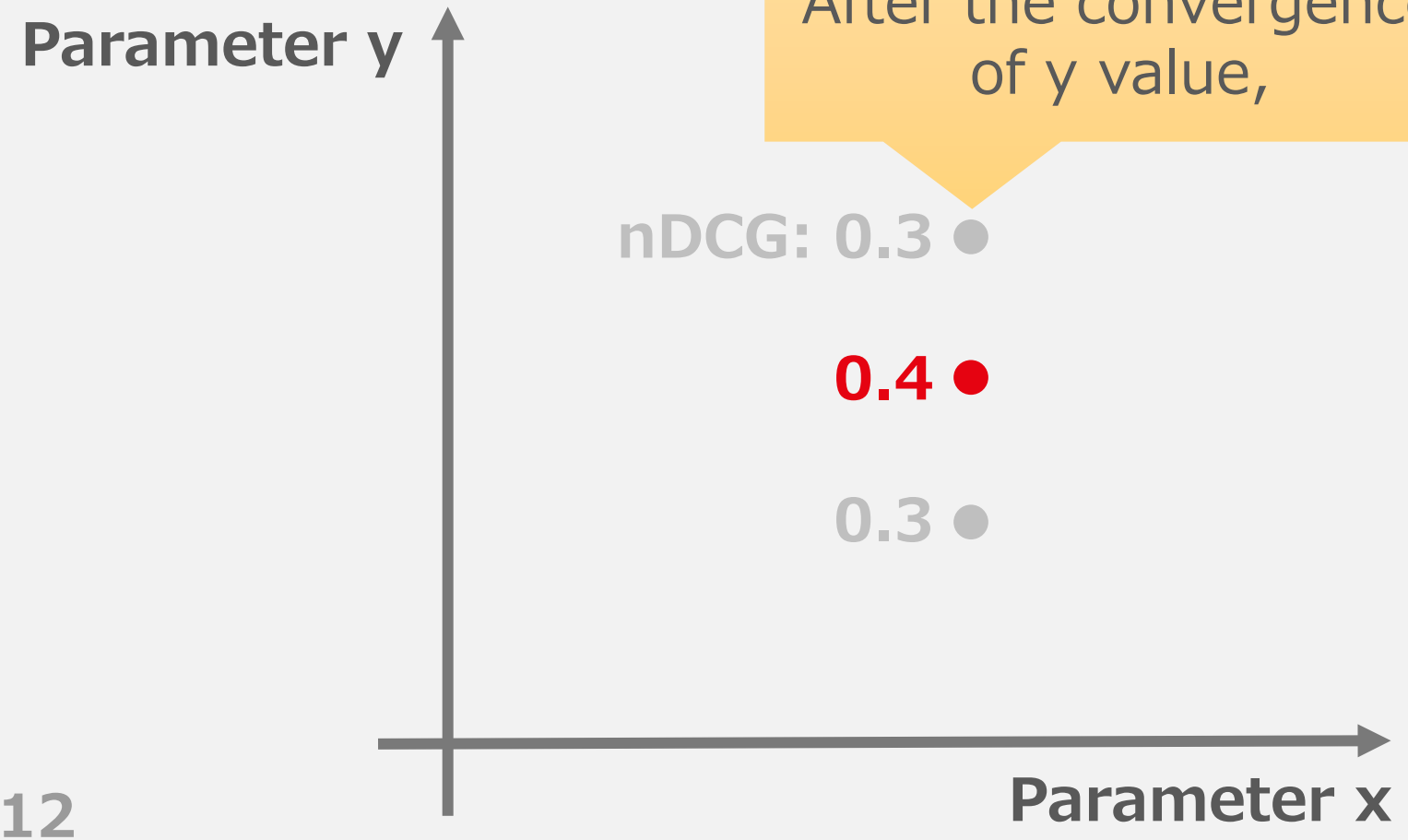
Baseline Method (cont'd)

CA optimizes one parameter at once.



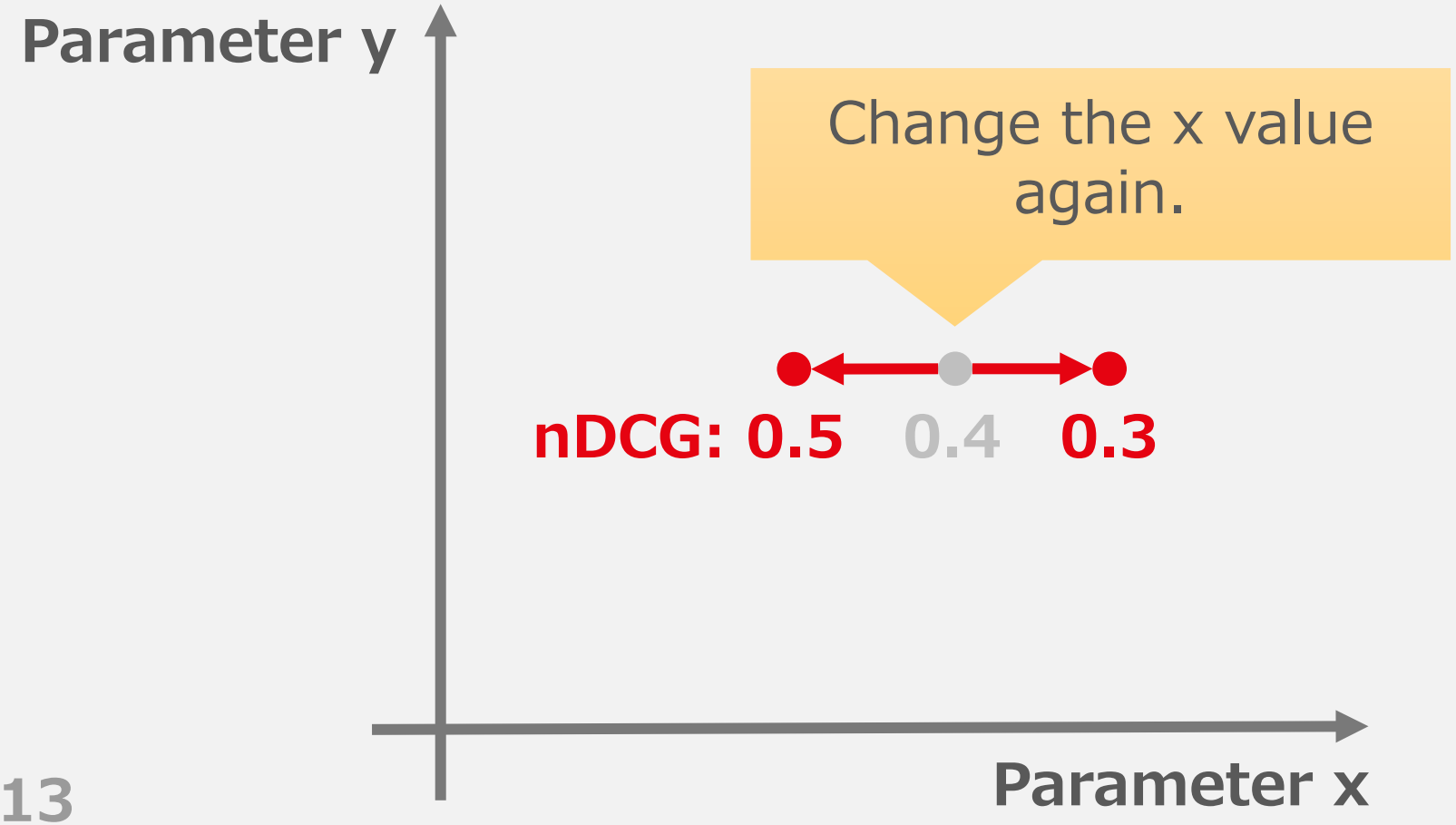
Baseline Method (cont'd)

CA optimizes one parameter at once.



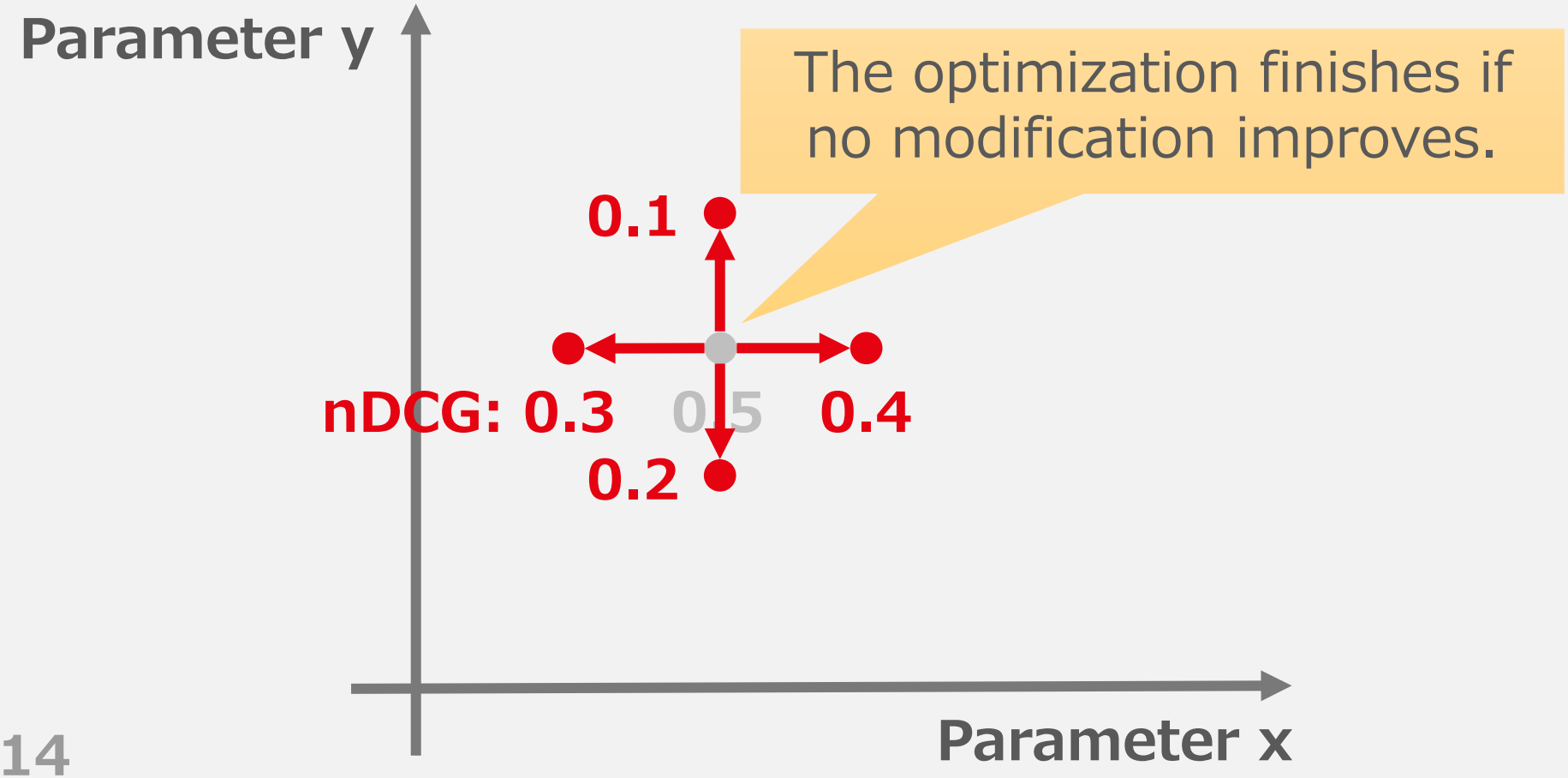
Baseline Method (cont'd)

CA optimizes one parameter at once.



Baseline Method (cont'd)

CA optimizes one parameter at once.



The BM25F Function

The BM25F Function

is a document scoring function which considers TFs on multiple fields,

- e.g. title, snippet, question, answer.

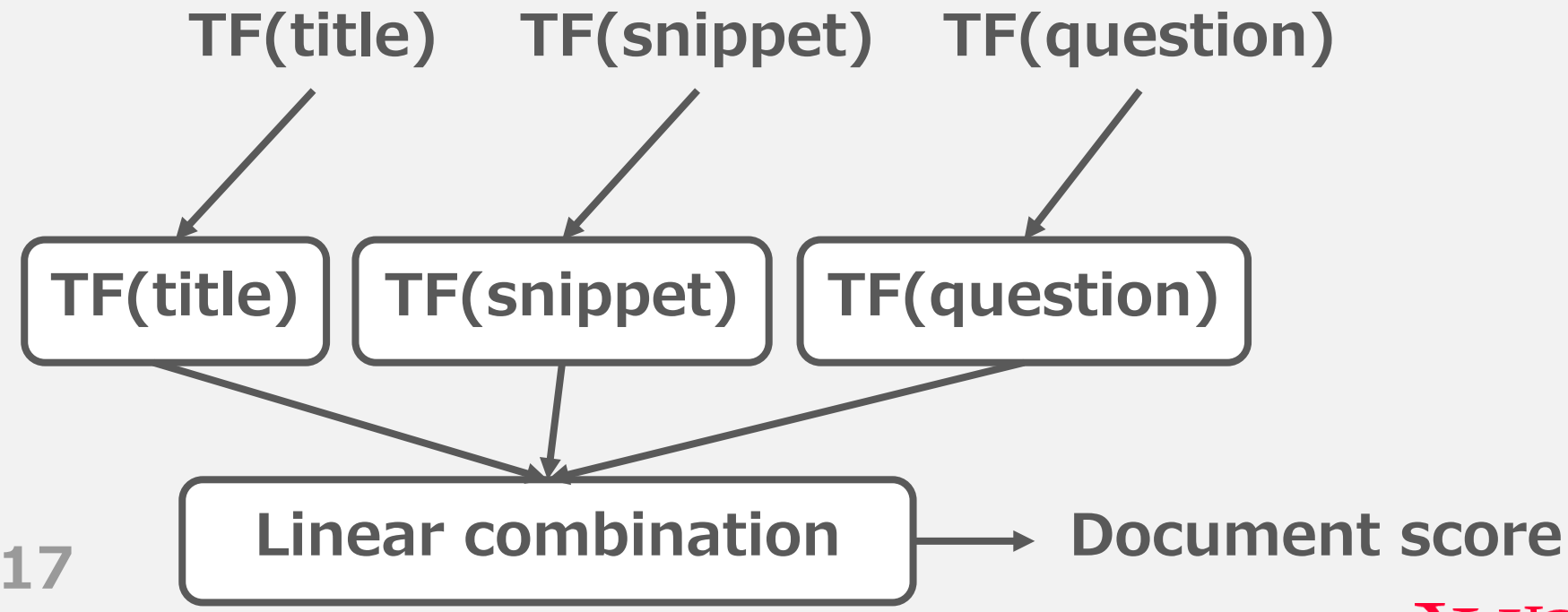
$$\sum_{t \in Q} \frac{w(t, D)}{k_1 + w(t, D)} \log \frac{N - \mathrm{df}(t) + 0.5}{\mathrm{df}(t) + 0.5},$$

$$w(t, D) = \sum_{f \in D} \frac{\mathrm{tf}(t, f, D) \cdot boost_f}{(1 - b_f) + b_f \cdot \mathrm{len}(f, D) / \mathrm{avgLen}(f)}$$

BM25F as Ranking Feature

The BM25F is a non-linear function.

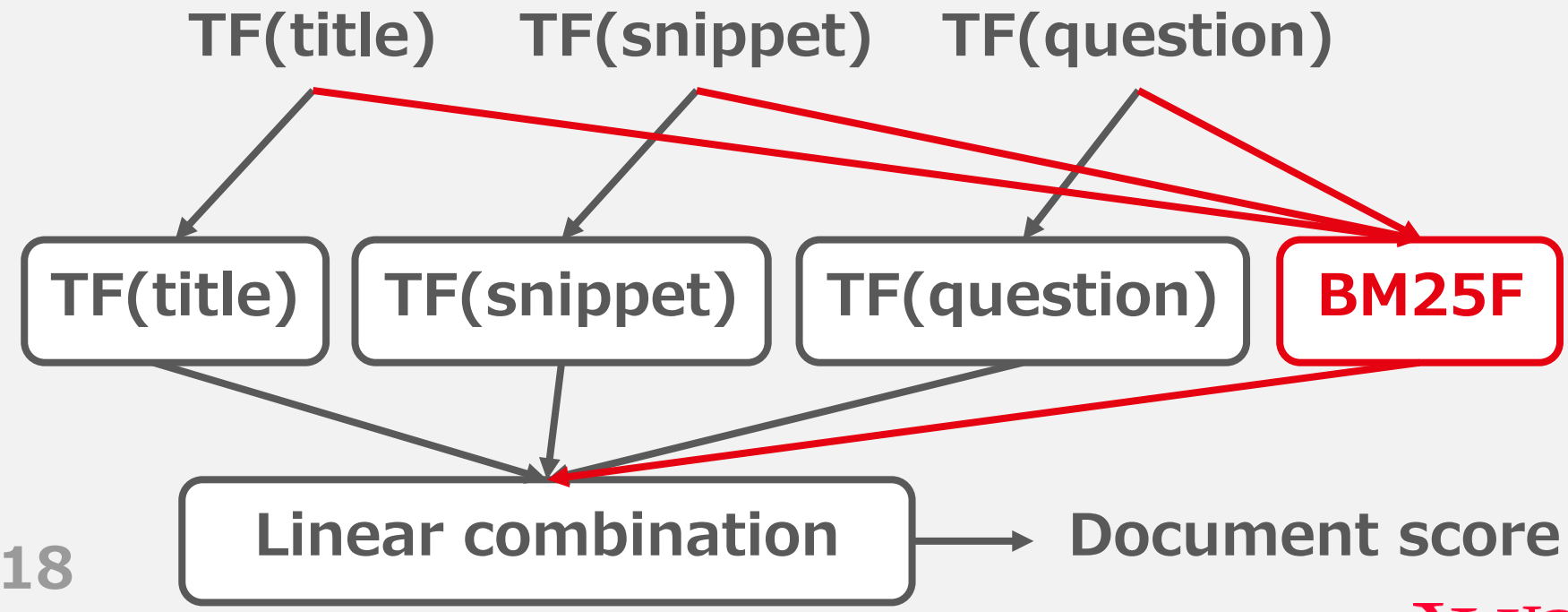
- Adding it to a linear combination may improve the model (cf. neural net.).



BM25F as Ranking Feature

The BM25F is a non-linear function.

- Adding it to a linear combination may improve the model (cf. neural net.).



Extended BM25F

We also added numeric field values to BM25F as well as TFs.

$$\sum_{t \in Q} \frac{w(t, D) + \alpha(D)}{k_1 + w(t, D) + \alpha(D)} \log \frac{N - \mathrm{df}(t) + 0.5}{\mathrm{df}(t) + 0.5}$$

$$\alpha(D) = \sum_{f \in D_N} v(f, D) \cdot boost_f$$

$$w(t, D) = \sum_{f \in D} \frac{\mathrm{tf}(t, f, D) \cdot boost_f}{(1 - b_f) + b_f \cdot \mathrm{len}(f, D) / \mathrm{avgLen}(f)}$$

YAHOO! JAPAN

BM25F as Ranking Feature

BM25F includes multiple parameters.

- We added three settings of BM25F to the ranking features.

Name	Considers
Naïve	All fields
SERP	Fields shown on SERPs
SERP+	Fields prominently shown on SERPs

- Field weights and other parameters were set by hand.

Summary of Our Runs

Run ID	nDCG	Description
5	0.344	Baseline
38	0.380	Baseline + naïve BM25F feature
50	0.412	5-fold CV
77	0.396	Baseline + three BM25F features
86	0.419	nDCG@10 as objective function

Summary of Our Runs

Run ID	nDCG	Description
5	0.344	Baseline
38	0.380	Baseline + naïve BM25F feature
50	0.412	5-fold CV
77	0.396	Baseline + three BM25F features
86	0.419	nDCG@10 as objective function

- Adding the naïve BM25F feature improved the nDCG@10 score +10%.

Cross Validation

- We applied traditional 5-fold cross validation to the model training.

Cross Validation

- We applied traditional 5-fold cross validation to the model training.

Run ID	nDCG	Description
5	0.344	Baseline
38	0.380	Baseline + naïve BM25F feature
50	0.412	5-fold CV
77	0.396	Baseline + three BM25F features
86	0.419	nDCG@10 as objective function

- The validation improved the nDCG@10 score +8.4%.

YAHOO! JAPAN

Objective Function

We changed objective function of CA.

Objective Function

We changed objective function of CA.

- First we used MAP instead of nDCG@10
- because quality of lower-ranked documents may be important in the greedy optimization.
  - E.g. a relevant document on 11th is more likely to be promoted to top-10 in future rather than that on 12th or lower.

Summary of Our Runs

Run ID	nDCG	Description
5	0.344	Baseline
38	0.380	Baseline + naïve BM25F feature
50	0.412	5-fold CV
77	0.396	Baseline + three BM25F features
86	0.419	nDCG@10 as objective function

- Finally, directly optimizing to nDCG@10 further improved the score (+5.7%).

Failed Attempts

Failed Attempts

- Optimizing BM25F parameters with CA
- Using more sophisticated models and learning methods:
  - Random Forests
  - LambdaMART

One possible explanation of the failure is due to the small size of training data.

Evaluation

Evaluation

- Feature importance
- Offline test results
  - Of our runs
  - Of our best run and the other teams' runs
- Online test results
  - Based on total credit
  - Based on win-loss ratio

Feature Importance

Feature Importance

For each feature, assigning 0 weight to it, we re-calculated nDCG@10.

- Lower score -> more importance

Rank	nDCG	Feature
1	0.193	# of PVs
2	0.2087	Log(# of answers)
3	0.2091	# of answers
4	0.210	NormTF(Snippets)
5	0.2111	Last modified date
6	0.2112	Length of title
7	0.2119	LM(Answer text)

Feature Importance

For each feature, setting 0 weight to it, we re-calculated nDCG@10.

- Lower score -> more importance

Rank	nDCG	Feature
1	0.193	# of PVs
2	0.2087	Log(# of answers)
3	0.2091	# of answers
4	0.210	NormTF(Snippets)
5	0.2111	Last modified date
6	0.2112	Length of title
7	0.2119	LM(Answer text)

Query-independent question *popularity* is most important

Feature Importance

For each feature, setting 0 weight to it, we re-calculated nDCG@10.

- Lower score -> more importance

Rank	nDCG	Feature
1	0.193	# of PVs
2	0.2087	Log(# of answers)
3	0.2091	# of answers
4	0.210	NormTF(Snippets)
5	0.2111	Last modified date
6	0.2112	Length of title
7	0.2119	LM(Answer text)

Matching between the query and fields on SERPs is also important

YAHOO! JAPAN

Feature Importance

For each feature, setting 0 weight to it, we re-calculated nDCG@10.

- Lower score -> more importance

Rank	nDCG	Feature
1	0.193	# of PVs
2	0.2087	Log(# of answers)
3	0.2091	# of answers
4	0.210	NormTF(Snippets)
5	0.2111	Last modified date
6	0.2112	Length of title
7	0.2119	LM(Answer text)

Other factors are information fresh-ness, amount, …

YAHOO! JAPAN

Feature Importance (cont'd)

Eventually, our BM25F features were more or less (but not so) important.

Rank	nDCG	Feature
…	…	…
25	0.2143	BM25F(SERP)
…	…	…
33	0.2144	BM25F(Naïve)
…	…	…
62	0.2148	BM25F(SERP+)
…	…	…
80	0.2149	TF(Question text)

YAHOO! JAPAN

Offline Test Results of Our Best Run and Other Teams' Runs

Offline Test Results of Our Best Run and Other Teams' Runs

Team	nDCG@10	ERR@10	Q-measure
OKSAT	.445	.276	.700*
YJRS	.419	.254	.713
cdlab	.418	.264	.697*
ORG	.413	.249	.702*
Erler	.406	.245	.707*
…	…	…	…

*: Statistically significantly different from the score of YJRS (paired t-test, $p < 0.05$)

YAHOO! JAPAN

Offline Test Results of Our Best Run and Other Teams' Runs

Team	nDCG@10	ERR@10	Q-measure
OKSAT	.445	.276	.700*
YJRS	.419	.254	.713
cdlab	.418	.264	.697*
ORG	.413	.249	.702*
Erler	.406	.245	.707*
..			

Our run achieved the 2nd nDCG, 3rd ERR, and best Q scores.

*: Sta
the score of YJRS (paired t-test, $p < 0.05$)

YAHOO! JAPAN

Offline Test Results of Our Best Run and Other Teams' Runs

Team	nDCG@10	ERR@10	Q-measure
OKSAT	.445	.276	.700*
YJRS	.419	.254	.713
cdlab	.418	.264	.697*
ORG	.413	.249	.702*
Erler	.406	.245	.707*

*: S
t)

Only Q was sensitive enough to indicate statistical significance, and the advantage of our run was significant.

YAHOO! JAPAN

Online Test Results based on Total Credit

Online Test Results based on Total Credit

Team	Total credit
Erler	22.35k
YJRS	22.31k
ORG	21.3k*
cdlab	20.0k*
Baseline (# of answers)	18.9k*
…	…

*: Statistically significantly different from the total credit of YJRS (paired t-test, $p < 0.05$)

Online Test Results based on Total Credit

Team	Total credit
Erler	22.35k
YJRS	22.31k
ORG	21.3k*
cdlab	20.0k*
Baseline (# of answers)	18.9k*

Our run obtained the 2nd-most amount of total credit.

*: Statist m the total c < 0.05)

YAHOO! JAPAN

Online Test Results based on Total Credit

Team	Total credit
Erler	22.35k
YJRS	22.31k
ORG	21.3k*
cdlab	20.0k*
Baseline (# of answers)	18.9k*

Difference from the 1st was not statistically significant.

*: Statist m the total c < 0.05)

YAHOO! JAPAN

Online Test Results based on Total Credit

Team	Total credit
Erler	22.35k
YJRS	22.31k
ORG	21.3k*
cdlab	20.0k*
Baseline (# of answers)	18.9k*
…	…

*: Sta[illegible] the tot[illegible].05)

Difference from the 3rd and the lower were statistically significant.

YAHOO! JAPAN

Online Test Results based on Win-Loss PV count

Online Test Results based on Win-Loss PV count

Opponent team	PVs we won	PVs we lost	Win-loss ratio
Erler	35.9k	30.8k	.538*
cdlab	40.5k	31.5k	.563*
ORG	37.0k	28.5k	.565*
Baseline (# of answers)	43.5k	24.7k	.637*
TUA1	46.1k	24.8k	.650*
…	…	…	…

*: Statistically significantly different from .500 (chi-square test, $p < 0.05$)

YAHOO! JAPAN

Online Test Results based on Win-Loss PV count

Opponent team	PVs we won	PVs we lost	Win-loss ratio
Erler	35.9k	30.8k	.538*
cdlab	40.5k	31.5k	.563*
ORG	37.0k	28.5k	.565*
Baseline (# of answers)	43.5k	24.7k	.637*
TUA1	46.1k	24.8k	.650*

Our run consistently achieved the win-loss ratios better than 0.5 against all the other runs with statistically significant differences of the PVs.

YAHOO! JAPAN

Conclusion

Conclusion

Our method performed well.

- This must be because of its simplicity and robustness.
- The BM25F is more or less useful as learning-to-rank features.
- Classical techniques are still useful.
  - Coordinate Ascent
  - Cross validation

EOP

YAHOO! JAPAN

Appendix

Table 1: Offline evaluation results of our runs.

ID	Description	nDCG@10
5	Test run.	0.34371
10	Naive BM25F.	0.36452
16	Roughly optimized BM25F.	0.33337
25	BM25F, roughly optimized with CA where $n = 3$.	0.33341
28	BM25F, roughly optimized with CA where $n = 3$ and $sf = 0.8$.	0.34316
38	Baseline + naive BM25F.	0.37965
48	Five-fold cross validation.	0.37965
50	Five-fold cross validation (fix).	0.41157
66	Five-fold cross validation (2).	0.40167
71	8foldCV_RandomForest	0.37091
77	Baseline + multiple BM25F features.	0.39637
82	8foldCV_LambdaMART	0.38087
86	Baseline + multiple BM25F features + nDCG@10.	**0.41894**

YAHOO! JAPAN

Table 4: Resulting p-values of Student's paired t-test among our runs in nDCG@10.

	5	10	16	25	28	38	48	50	66	71	77	82	86
5		0.0641	0.3639	0.4017	0.9679	0.0183	0.0183	0.0000	0.0004	0.1053	0.0011	0.0260	0.0000
10	0.0641		0.0111	0.0103	0.0912	0.2722	0.2722	0.0008	0.0136	0.7190	0.0225	0.3004	0.0002
16	0.3639	0.0111		0.9906	0.1870	0.0004	0.0004	0.0000	0.0000	0.0117	0.0000	0.0003	0.0000
25	0.4017	0.0103	0.9906		0.1228	0.0005	0.0005	0.0000	0.0000	0.0125	0.0000	0.0003	0.0000
28	0.9679	0.0912	0.1870	0.1228		0.0047	0.0047	0.0000	0.0000	0.0768	0.0001	0.0077	0.0000
38	0.0183	0.2722	0.0004	0.0005	0.0047			0.0001	0.0042	0.5323	0.0480	0.9200	0.0001
48	0.0183	0.2722	0.0004	0.0005	0.0047			0.0001	0.0042	0.5323	0.0480	0.9200	0.0001
50	0.0000	0.0008	0.0000	0.0000	0.0000	0.0001	0.0001		0.1417	0.0051	0.0037	0.0181	0.2075
66	0.0004	0.0136	0.0000	0.0000	0.0000	0.0042	0.0042	0.1417		0.0245	0.4406	0.0935	0.0372
71	0.1053	0.7190	0.0117	0.0125	0.0768	0.5323	0.5323	0.0051	0.0245		0.0909	0.2822	0.0013
77	0.0011	0.0225	0.0000	0.0000	0.0001	0.0480	0.0480	0.0037	0.4406	0.0909		0.2534	0.0007
82	0.0260	0.3004	0.0003	0.0003	0.0077	0.9200	0.9200	0.0181	0.0935	0.2822	0.2534		0.0052
86	0.0000	0.0002	0.0000	0.0000	0.0000	0.0001	0.0001	0.2075	0.0372	0.0013	0.0007	0.0052	

Table 5: The best run of all teams in nDCG@10, and their corresponding ERR@10 and Q-measure.

ID	Team	nDCG@10	ERR@10	Q-measure
7	ORG	0.41328	0.24942	0.70247
18	KUIDL	0.35788	0.21967	0.67360
19	TUA1	0.37670	0.23338	0.69432
22	Erler	0.40566	0.24507	0.70657
54	SLOLQ	0.31908	0.19760	0.65563
83	cdlab	0.41800	0.26381	0.69732
86	YJRS	0.41894	0.25391	0.71339
88	OKSAT	0.44471	0.27605	0.69980

Table 6: Resulting p-values of Student's paired t-test between our run and each of other teams' runs.

ID	Team	nDCG@10	ERR@10	Q-measure
7	ORG	0.50645	0.67545	0.00002
18	KUIDL	0.00003	0.07254	0.00000
19	TUA1	0.00052	0.13913	0.00003
22	Erler	0.07422	0.39101	0.00000
54	SLOLQ	0.00000	0.00233	0.00000
83	cdlab	0.94031	0.53139	0.00000
88	OKSAT	0.09991	0.16356	0.00616

Table 7: All runs, their total credit in online test and win-loss counts of our run against each run.

ID	Team	Total credit	Win PV	Loss PV
7	ORG	21301.1	37010	28496
18	KUIDL	16935.8	47498	24552
19	TUA1	17285.2	46083	24772
22	Erler	22345.7	35912	30779
54	SLOLQ	14892.0	50273	20984
83	cdlab	19961.9	40529	31465
86	YJRS	22307.6		
88	OKSAT	16597.7	46958	25169
-	AS-IS	14037.1	52736	19832
-	N-ANS	18917.5	43452	24747

Table 8: Resulting p-values of Student's paired t-test for total credit and Pearson's chi-square test for win-loss counts of our best run against other runs.

ID	Team	Total credit	Win-loss PV
7	ORG	0.00000	0.00000
18	KUIDL	0.00000	0.00000
19	TUA1	0.00000	0.00000
22	Erler	0.90975	0.00000
54	SLOLQ	0.00000	0.00000
83	cdlab	0.00000	0.00000
88	OKSAT	0.00000	0.00000
-	AS-IS	0.00000	0.00000
-	N-ANS	0.00000	0.00000