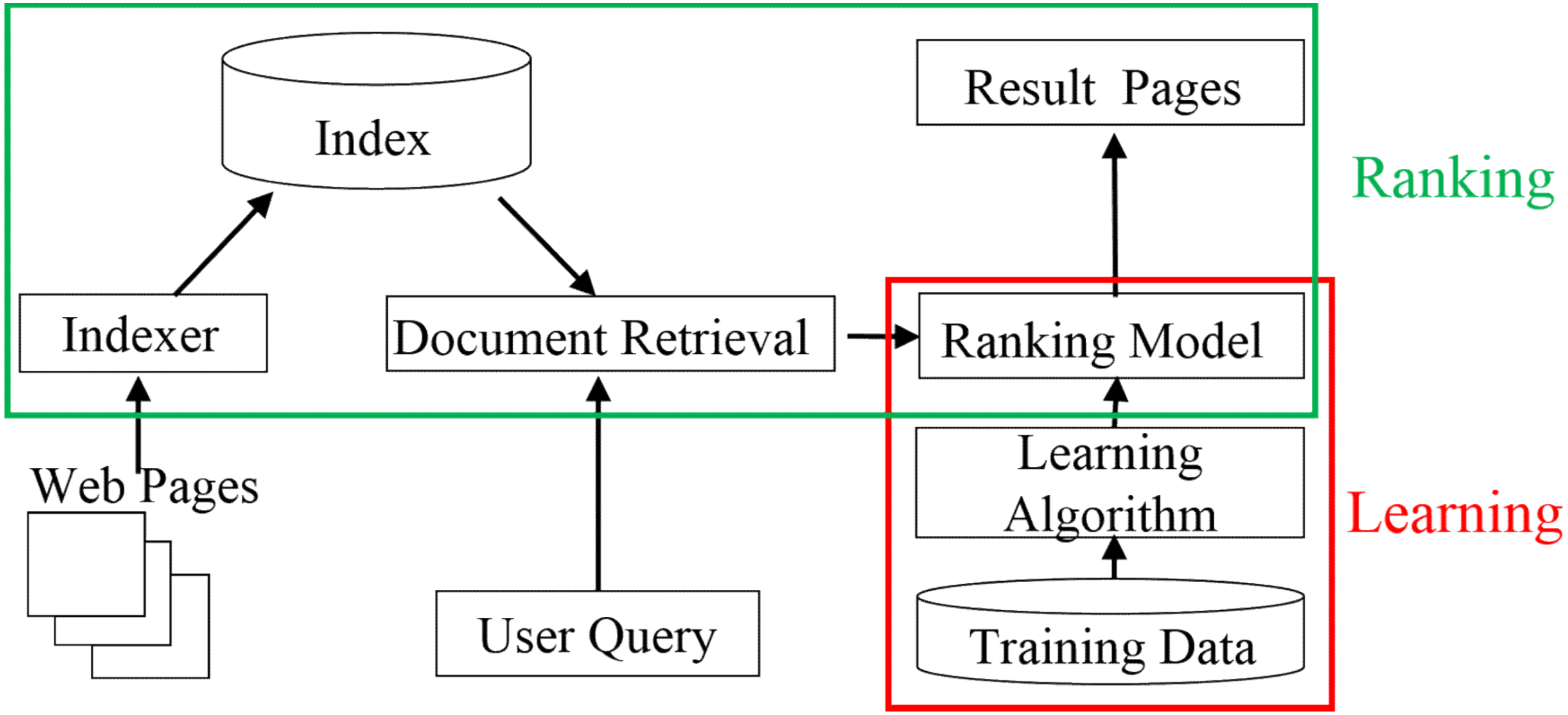
Imperial College
London

Accelerating Position-Aware Top-*k* ListNet for Ranking under Custom Precision Regimes

Qiang Li, Erwei Wang, Shane T. Fleming, David B. Thomas, Peter Y. K. Cheung

Imperial College London

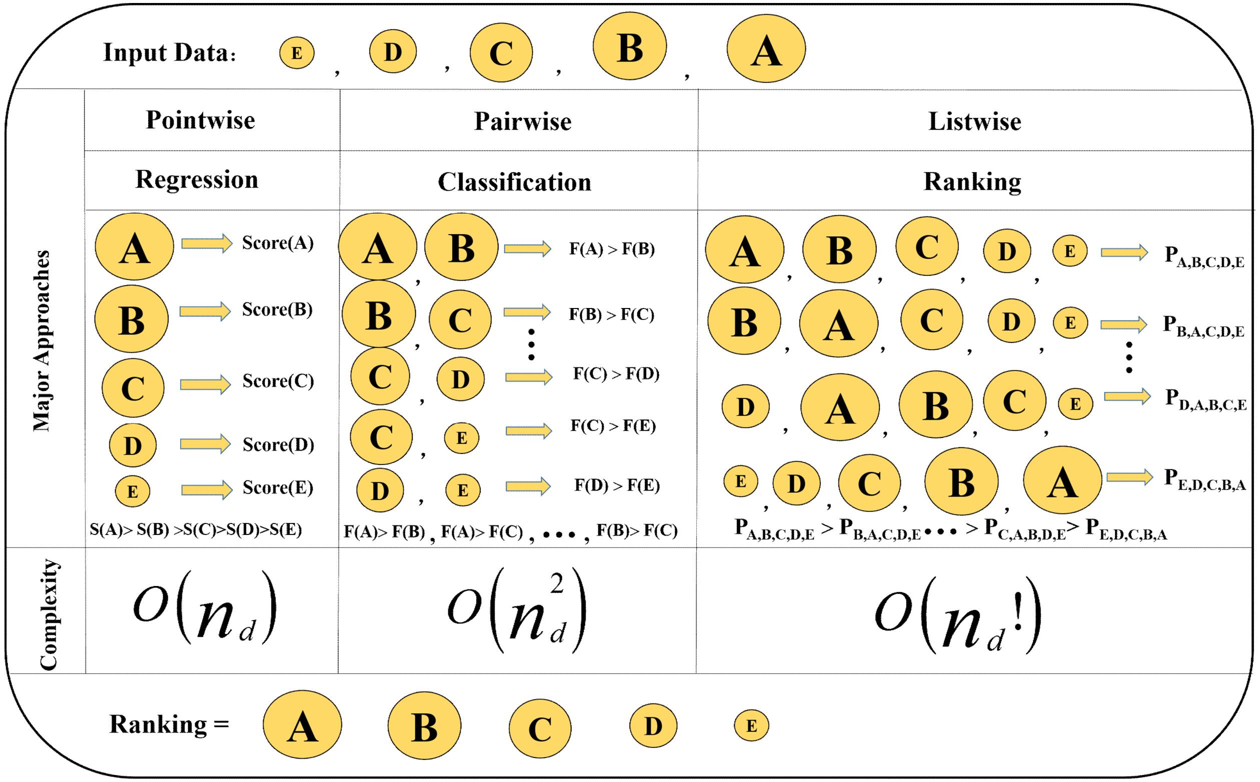
Machine-learning based search engine



Ranking plays a key role in information retrieval

Imperial College
London

Major Approaches for Ranking



Input Data: E, D, C, B, A

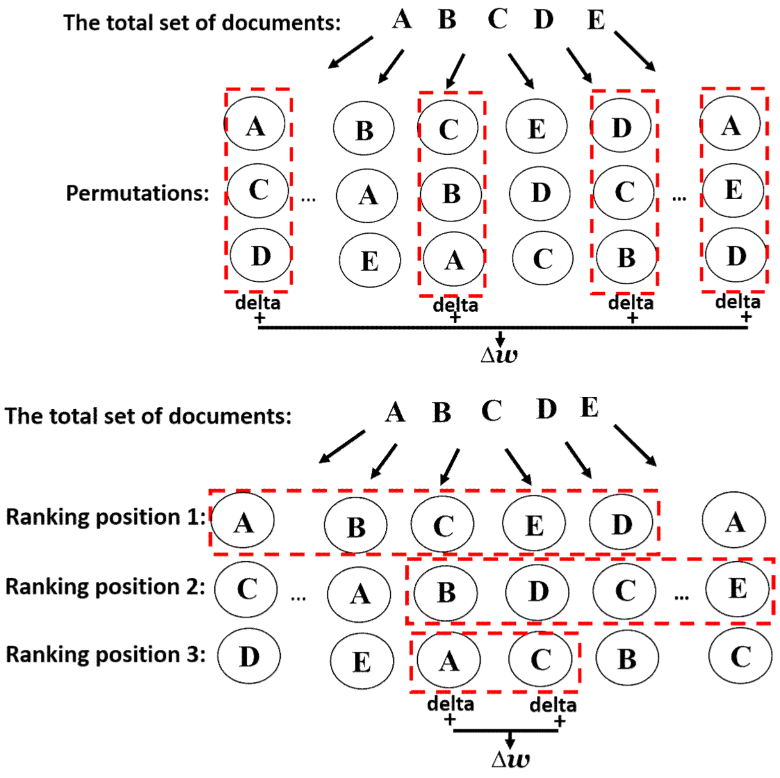
Major Approaches

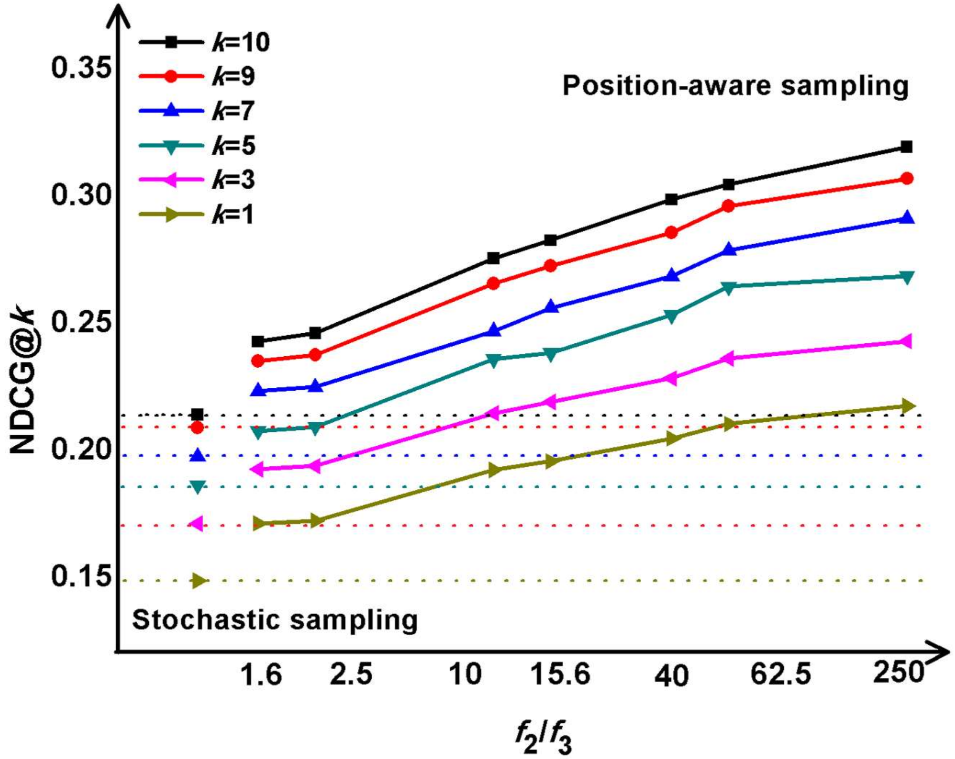
Pointwise	Pairwise	Listwise
Regression	Classification	Ranking
A → Score(A) B → Score(B) C → Score(C) D → Score(D) E → Score(E)	A, B → F(A) > F(B) B, C → F(B) > F(C) ⋮ C, D → F(C) > F(D) C, E → F(C) > F(E) D, E → F(D) > F(E)	A, B, C, D, E → $P_{A,B,C,D,E}$ B, A, C, D, E → $P_{B,A,C,D,E}$ ⋮ D, A, B, C, E → $P_{D,A,B,C,E}$ E, D, C, B, A → $P_{E,D,C,B,A}$
S(A)> S(B) >S(C)>S(D)>S(E)	F(A)> F(B) , F(A)> F(C) , • • • , F(B)> F(C)	$P_{A,B,C,D,E} > P_{B,A,C,D,E} \bullet\bullet\bullet > P_{C,A,B,D,E} > P_{E,D,C,B,A}$
Complexity: $O(n_d)$	$O(n_d^2)$	$O(n_d!)$

Ranking = A B C D E

Challenge of listwise approach: high computational complexity

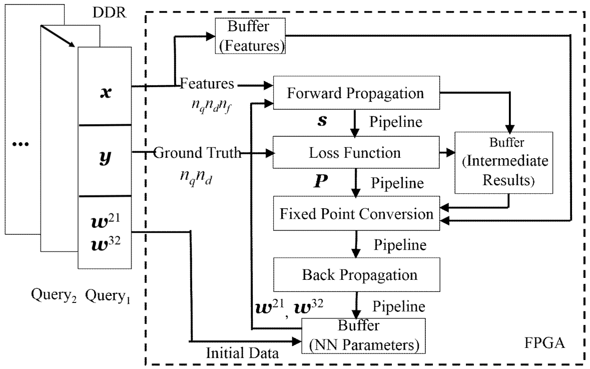
Position-Aware Sampling

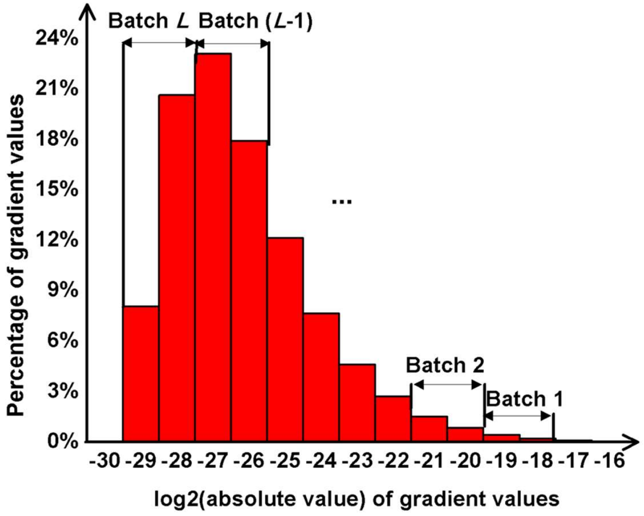


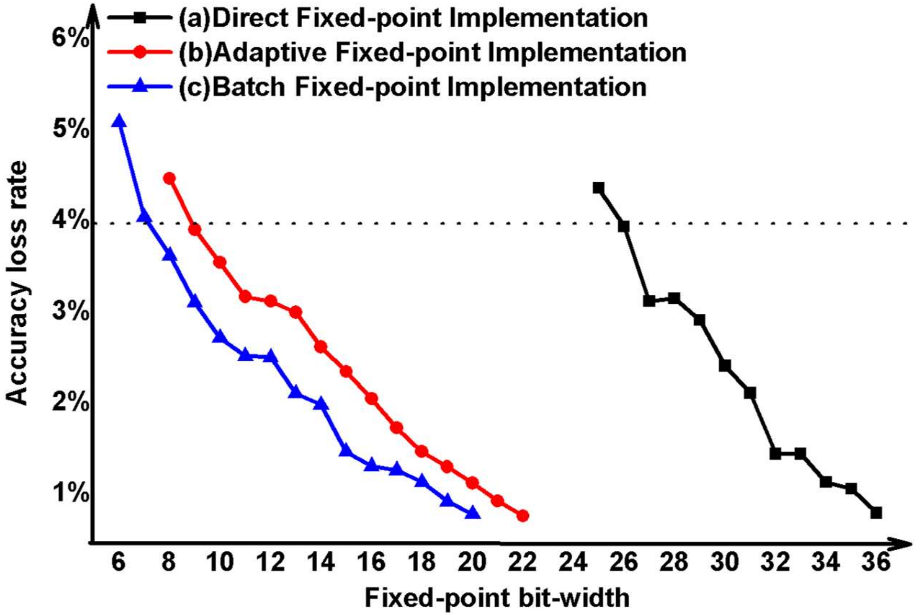


position-aware approach delivers a better accuracy than traditional stochastic sampling method

Batch Quantization







4.42x speedup over an Nvidia GTX 1080T GPU implementation with **2%** accuracy loss

Imperial College
London

Thank you!