

Online Learning from Click Data

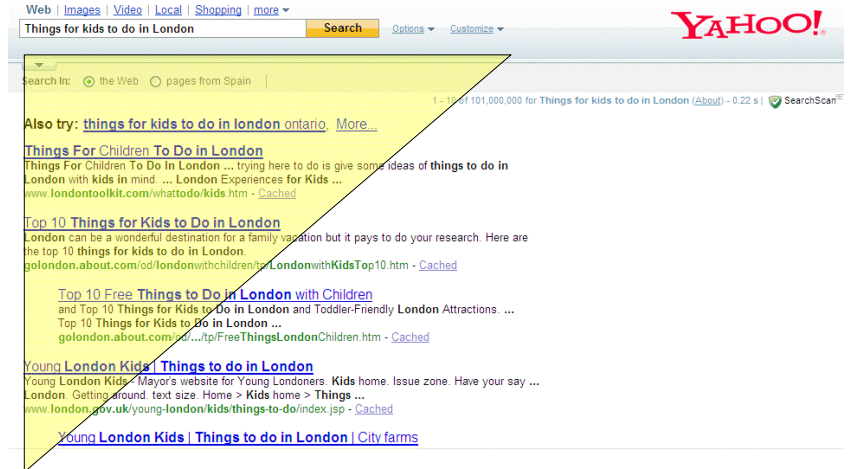
Vanessa Murdock

Ads: Massi Ciaramita, Vassilis Plachouras

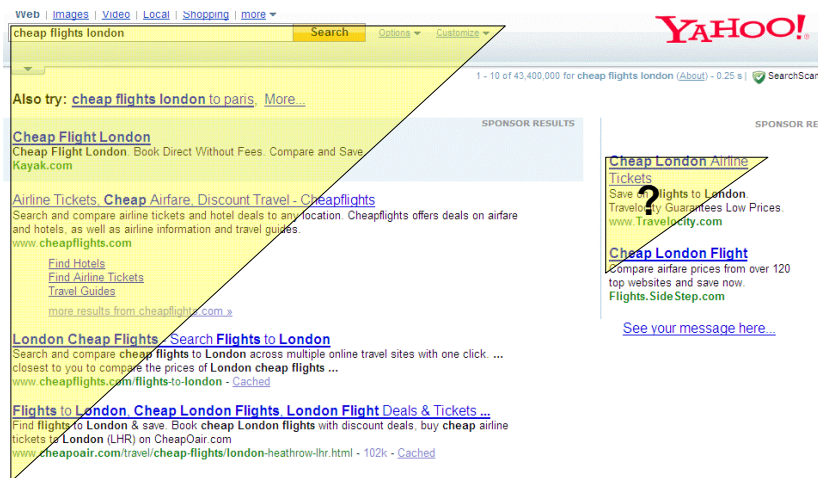
Images: Lluís Garcia, Ximena Olivares, Roelof van Zwol

Yahoo! Research Barcelona

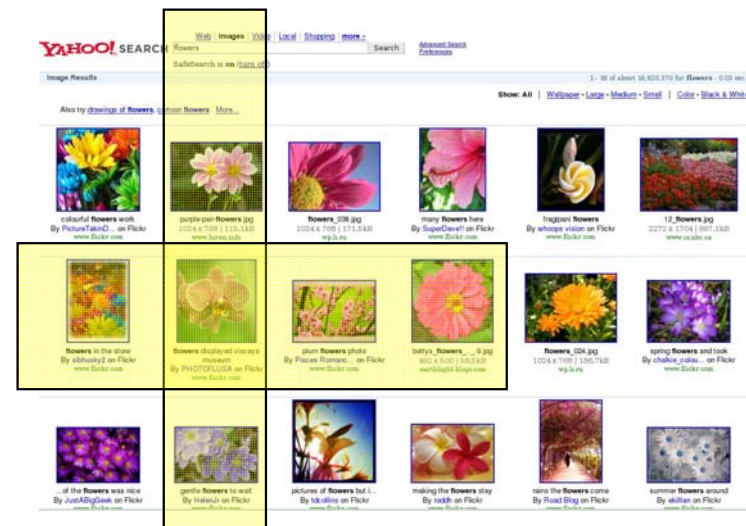
What Users Look At



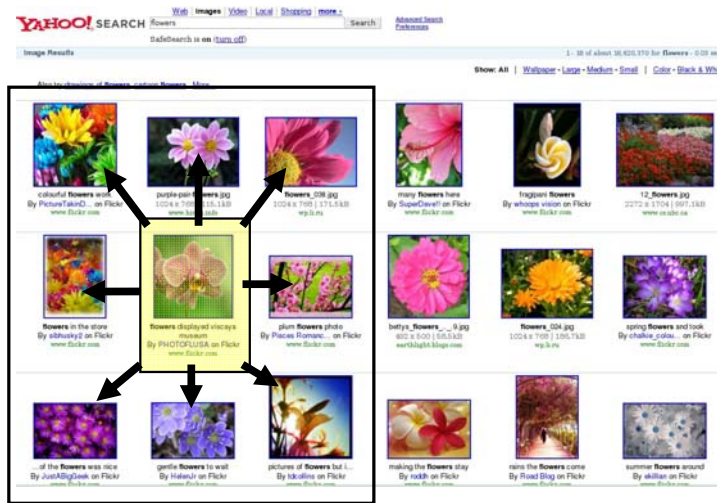
What about Ads?



...and Images?

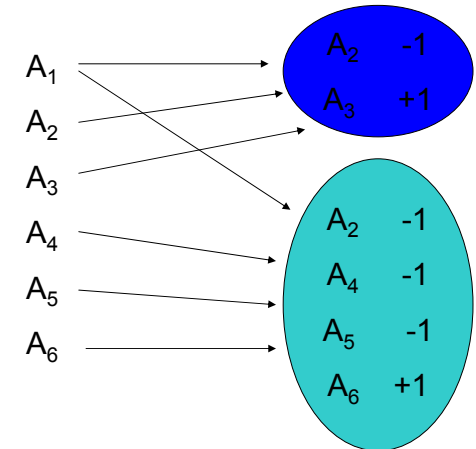


...and Images?



Learning from Clicks

- User clicks give relative preference
- Clicks at rank 1 ignored
- Train and evaluate in “blocks”
- 123,798 blocks



Learning Frameworks

- Perceptron
 - Results comparable to SVM
 - More efficient
- Online learning
 - Each pattern considered in isolation
 - Data need not be stored
- Task: Produce a ranking of ads given a query

Classification

- Two classes: clicked and nonclicked
 - Assume they are separable by a hyperplane
- Train on patterns independently
- Binary perceptron
 - Averaging: Average weight vector of all models posited during training
 - Uneven margin:
 - Clicked class outnumbered by nonclicked class
- Perceptron produces a score
 - Use the score to rank ads in each block

Ranking

- Train on pairs of patterns
- Only two possible ranks:
 - Clicked event given rank 1
 - Nonclicked events given rank 2
- Error function depends on pairwise scores
- Regularized by averaging

Multilayer Regression

- Nonlinear activation layer between input and output layers
- Sigmoidal nonlinear layer generates arbitrarily complex patterns
- Fully connected three-layer network
- Parameters initialized at random

Features, Part I

- **B**aseline: Cosine similarity
 - Ad materials concatenated
- **O**verlap features
 - Binary
 - All, some or none of query terms in ads
 - Percentage query terms in ads
- **F**ields
 - Cosine similarity with title, bidded phrases, description

Features, Part II

- **P**ointwise Mutual Information
 - Term pairs in patterns computed in query logs
 - Average PMI
 - Max PMI
- **C**hi-Squared Statistic
 - Term pairs in patterns computed in query logs
 - Number of query term-ad term pairs in top 5% of chi-squared statistics
- **B**ias feature: Value 1 for every pattern

Keyword Extraction

- Divergence from Randomness

Probability that t appears in the set of target documents

$$w(t) = tf_{t,p} \log \frac{1 + P_n}{P_n} + \log_2(1 + P_n)$$

Frequency of t in document p

Pointwise Mutual Information

- PMI of all possible keyword pairs
- Two Features
 - Average PMI
 - Maximum PMI
- Three corpora
 - The Web
 - UK2006 summary collection
 - Query Log

$$PMI(t_1, t_2) = \log \frac{P(t_1, t_2)}{P(t_1)P(t_2)}$$

Pearson's χ^2

- Compute χ^2 for pairs of keywords
- Percentile Rank method
 - χ^2 statistic compared to χ distribution
 - Comparison not reliable (too small magnitude)
 - Consider the percentile rank of the pair
- CSQ_z number of keyword pairs that have χ^2 in the top z% of all pairs

Experimental Setup

- 5 Development sets, 5 Test sets
 - Parameters set on development sets
 - Average 1500 blocks per set
- Training set had 109,000 blocks
- Models converged in an average of 10 iterations
- Multilayer parameters set to standard default values

Classification Results

	Prec @ 1	MRR
B	0.322	0.582
B+O	0.319	0.578
B+F	0.341	0.593
B+F+O	0.357	0.605
B+F+O+P	0.357	0.604
B+F+O+C	0.351	0.601
B+F+O+C+P	0.360	0.606

Ranking Results

	Prec @ 1	MRR
B	0.333	0.590
B+O	0.352	0.602
B+F	0.347	0.597
B+F+O	0.357	0.605
B+F+O+P	0.359	0.606
B+F+O+C	0.364	0.610
B+F+O+C+P	0.363	0.609

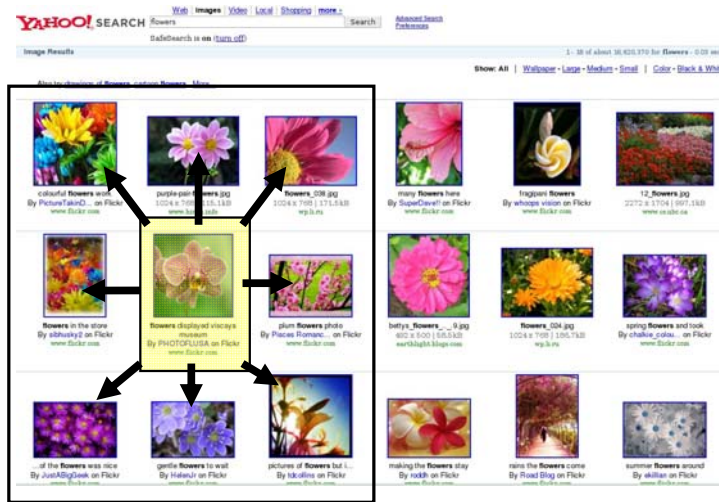
Regression Results

	Prec @ 1	MRR
B	0.328	0.585
B+O	0.343	0.596
B+F	0.374	0.615
B+F+O	0.371	0.614
B+F+O+P	0.374	0.617
B+F+O+C	0.381	0.619
B+F+O+C+P	0.388	0.624

Discussion

- PMI and Chi-squared features strongly correlated
 - Linear model lacks discriminative power to use information from both
 - Best strategy to trust one consistently
- Ranker outperforms classifier
 - Pairwise comparisons
- Reranking ads shown by Yahoo!
 - Results biased by initial retrieval

Images



Images

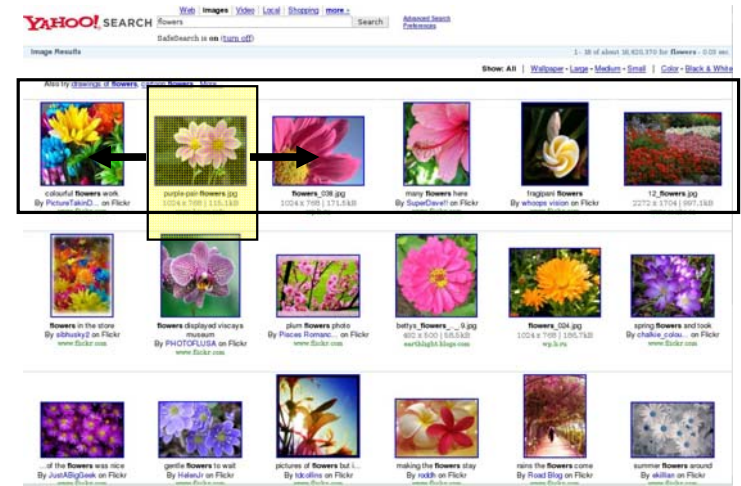


Image Results

	Prec @ 1	MRR
Baseline	0.4198	0.6186
Learned Baseline	0.4073	0.6104
Text Features	0.5484	0.7034

Thank You!