Here or There
Preference Judgments for Relevance

Carterette, Bennett, Chickering and Dumais

CS 6784 Advanced Machine Learning

Tobias Schnabel

Megan Baker

March 6, 2014

Motivation

- How much would you pay for each of these cars?

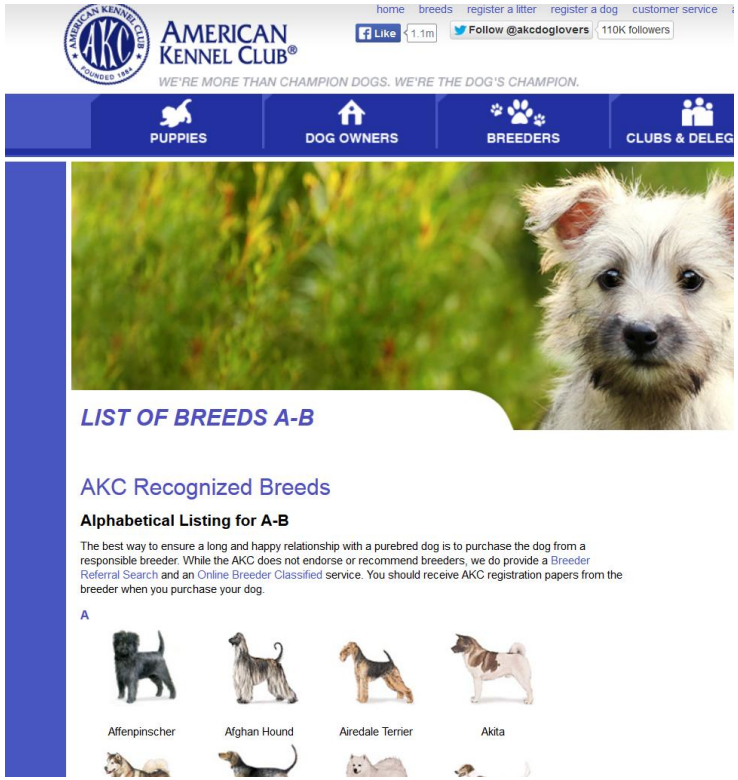
A. $30,000
B. $45,000
C. $90,000
D. $100,000

Motivation

- Now: For which car would you pay more?

Motivation

- Query: Dog breeds

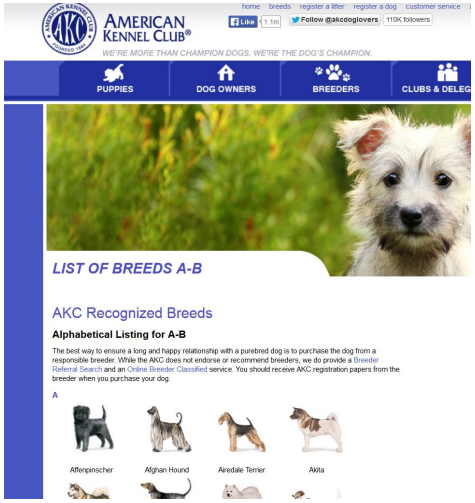


Fair | Good | Excellent | Perfect

Motivation

- How are search results evaluated?
  - Binary (relevant/non-relevant)
  - On a numeric scale (highly relevant .... less relevant)

- Problems:



Assessor study

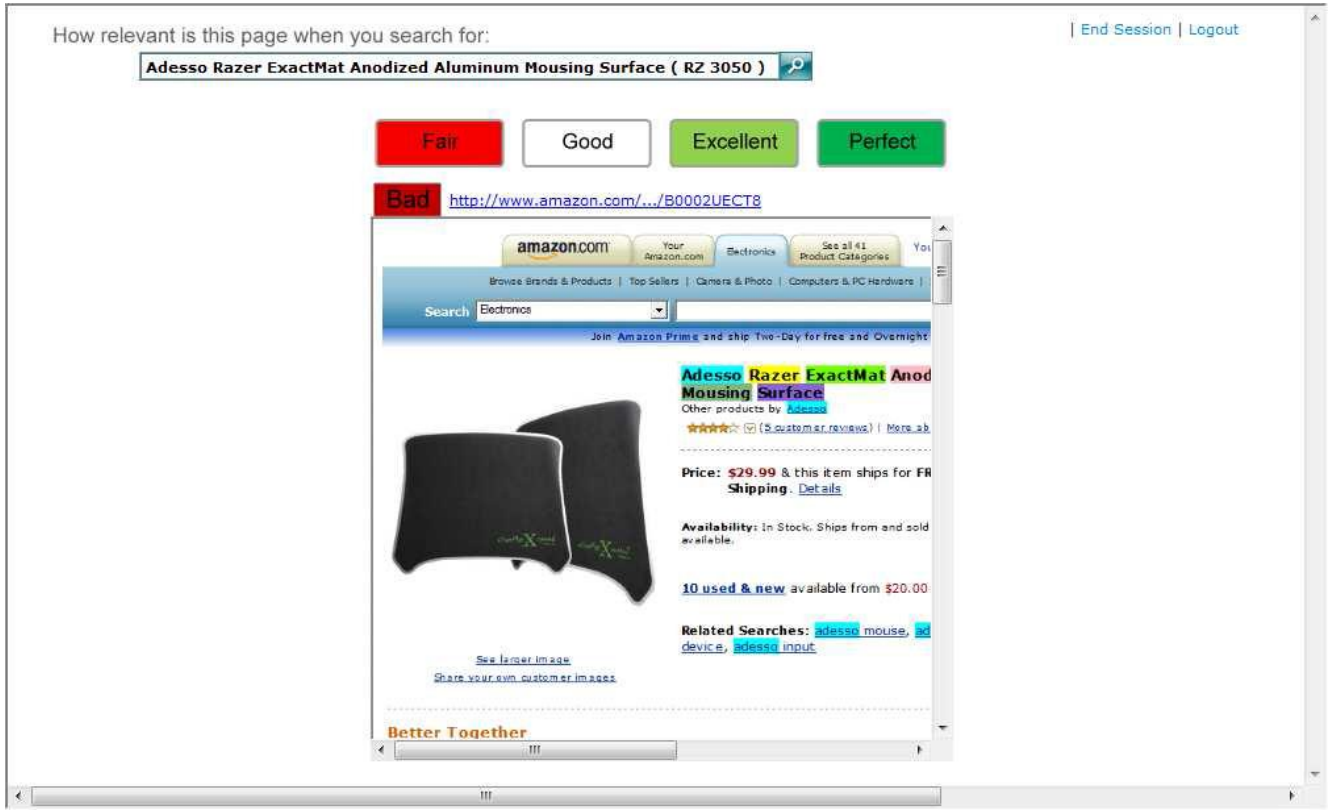
- Compared three types of judgments:
  - Absolute

  - Binary preferences

  - Graded preferences

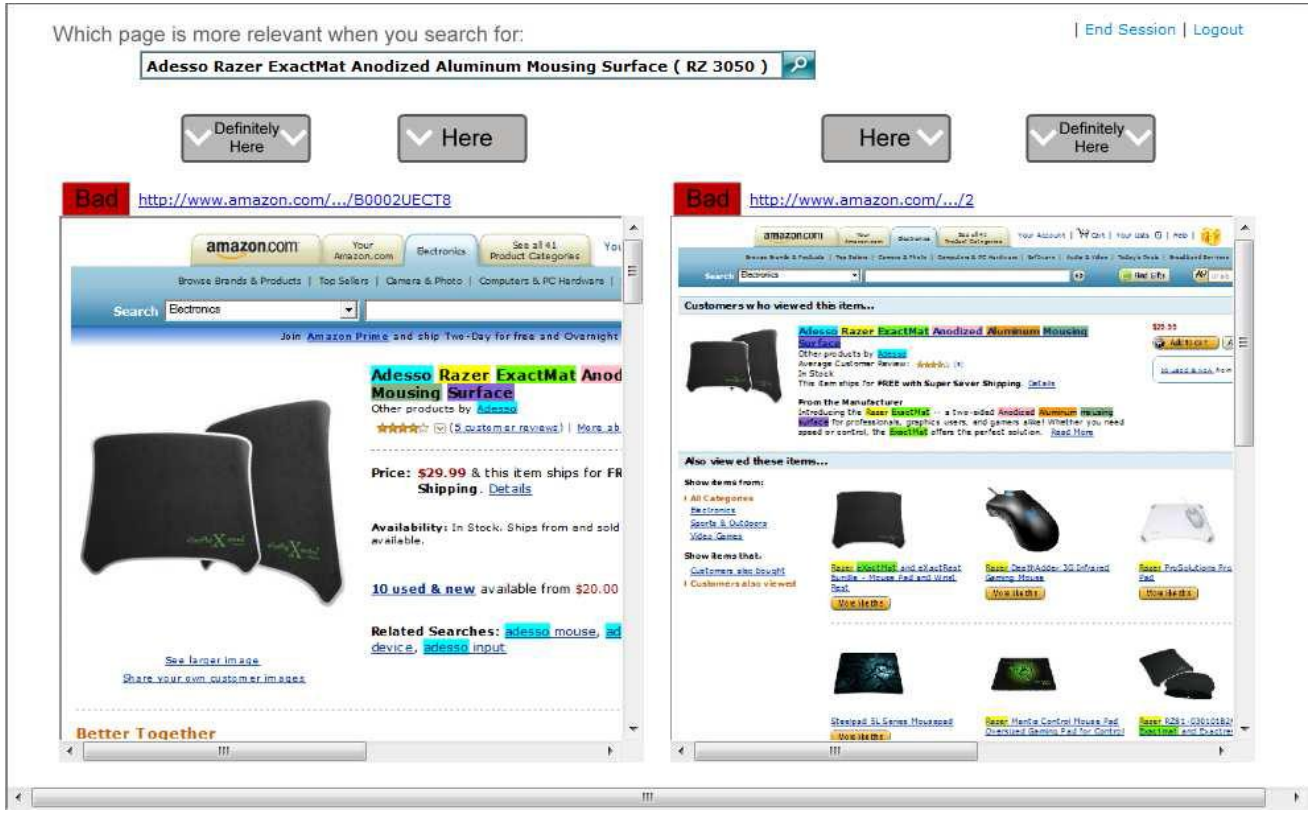
Assessor study

- Absolute



Assessor study

- Graded preferences



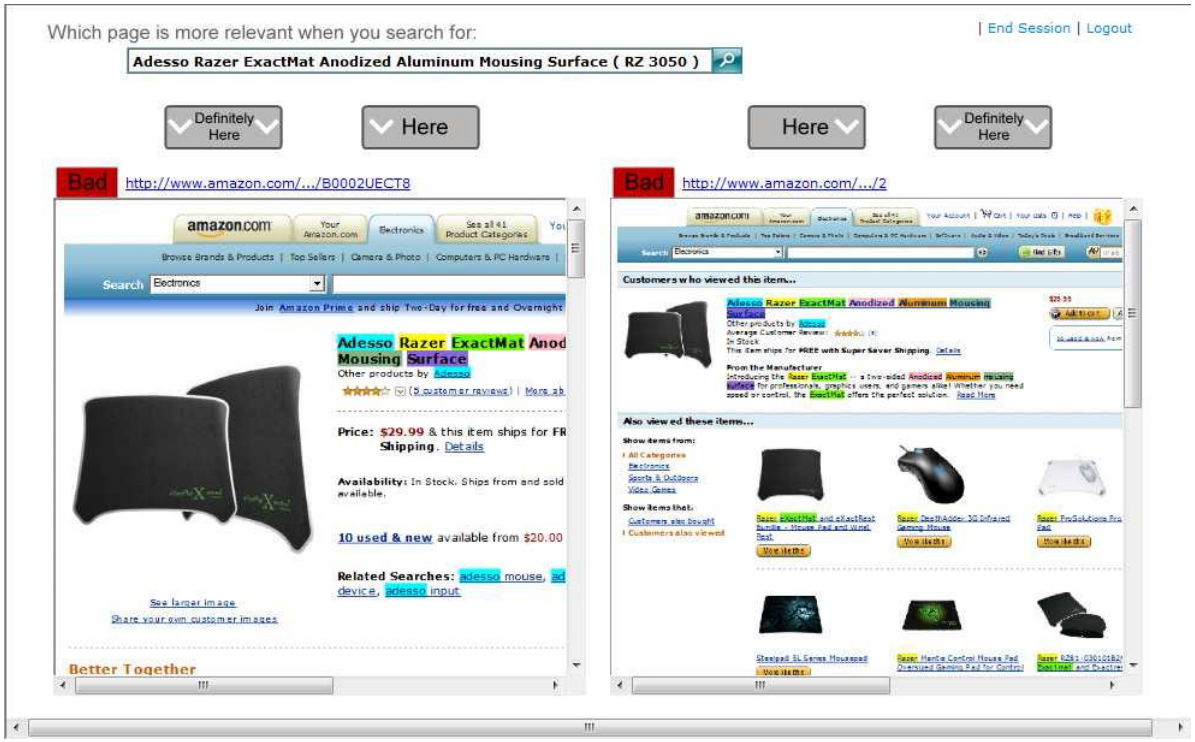
Assessor study

- Experimental set-up:
  - 51 queries, each ~12 results
  - Each <result, result> pair was rated for preferences

- Judgments
  - 6 assessors, all Microsoft employees
  - Guidelines for absolute five-point scale
  - Pages judged as „bad“ removed from results

Assessor study

- Transitivity:
  - Holds 99 % on average
  - But ...



Assessor study

- Low agreement for absolute judgements

	Bad	Fair	Good	Excellent	Perfect
Bad	0.579	0.29	0.118	0.014	0.000
Fair	0.208	0.332	0.309	0.147	0.003
Good	0.095	0.348	0.286	0.26	0.011
Excellent	0.011	0.167	0.264	0.535	0.022
Perfect	0.000	0.042	0.125	0.25	0.583

Assessor study

- Explicit preferences better than inferred from absolute statements

	A < B	A,B bad	A > B	Total
A < B	0.752	0.033	0.215	2580
A,B bad	0.208	0.567	0.225	413
A > B	0.201	0.034	0.765	2757

(a) explicit preferences

	A < B	A,B bad	A > B	Total
A < B	0.657	0.051	0.292	2530
A,B bad	0.297	0.38	0.323	437
A > B	0.278	0.053	0.669	2654

(b) inferred preferences

Assessor study

- Graded preferences vs. binary preferences

	A < B	A,B bad	A > B	Total
A < B	0.752	0.033	0.215	2580
A,B bad	0.208	0.567	0.225	413
A > B	0.201	0.034	0.765	2757

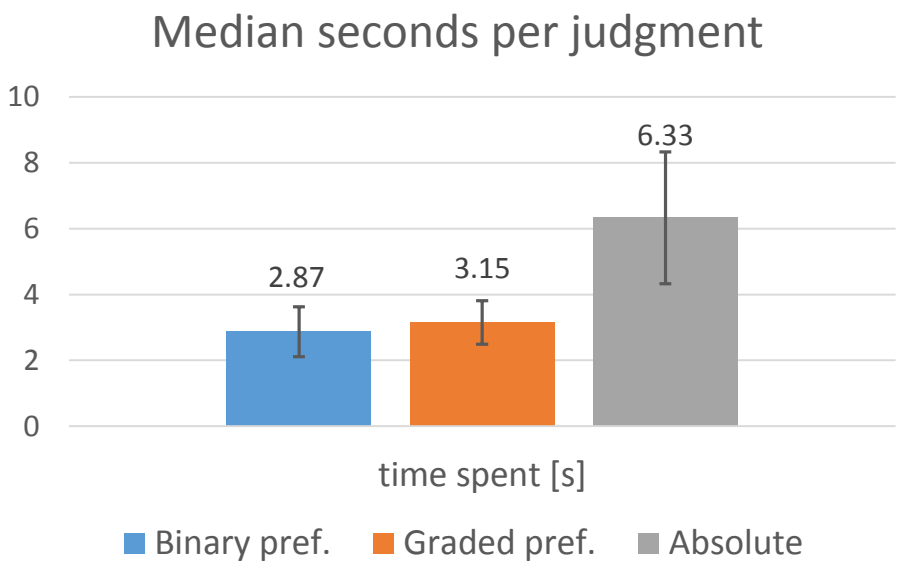
(a) preferences

	A << B	A < B	A, B bad	A > B	A >> B	Total
A << B	0.247	0.621	0.000	0.132	0.000	219
A < B	0.059	0.661	0.043	0.221	0.015	2288
A, B bad	0.000	0.244	0.453	0.300	0.002	406
A > B	0.012	0.212	0.051	0.670	0.055	2389
A >> B	0.000	0.180	0.005	0.680	0.134	194

(b) graded preferences

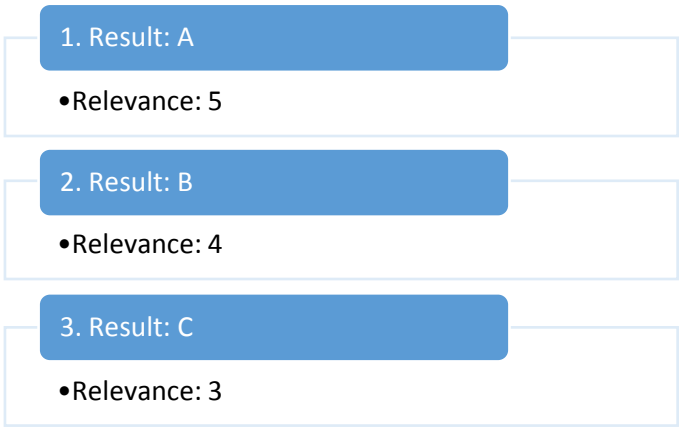
Assessor study

- Making preference statements is faster than making absolute statements

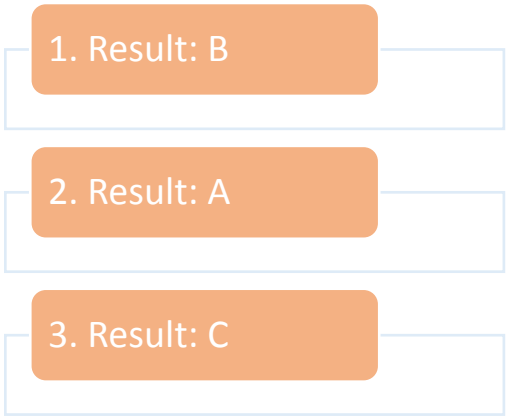


Evaluation measures

- Want to evaluate entire queries
- Traditionally: graded relevance



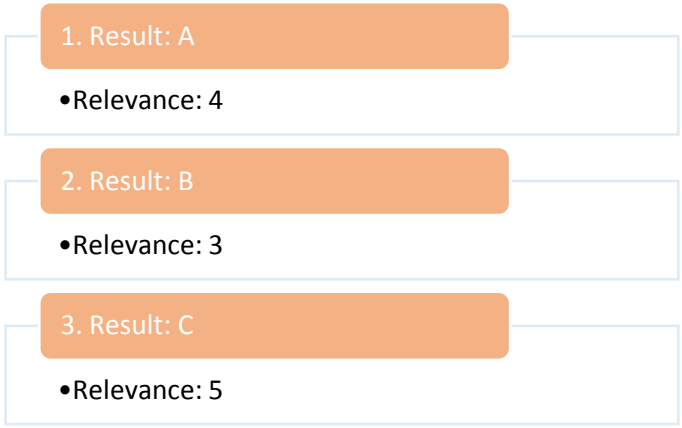
Gold standard from assessors



Search engine

Evaluation measures

- DCG (discounted cumulative gain)
  - DCG@k = $\sum_{i=1}^{k} \frac{2^{rel_i - 1}}{\log_2 i+1}$
  - nDCG@k = $\frac{\text{DCG@k}}{\text{iDCG@k}}$



Search engine

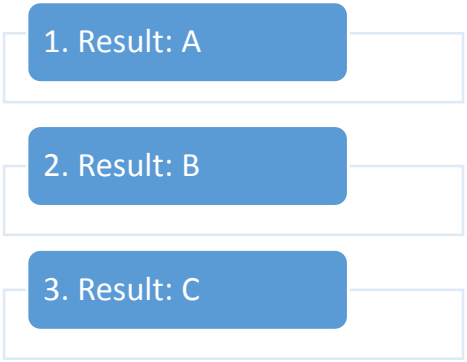
$$\frac{2^3}{\log_2 2} + \frac{2^2}{\log_2 3} + \frac{2^4}{\log_2 4}$$ = DCG@3

$$\frac{2^4}{\log_2 2} + \frac{2^3}{\log_2 3} + \frac{2^2}{\log_2 4}$$ = iDCG@3

Evaluation measures

- Activity: How would you evaluate preferences?

Assessor preferences:
A > B
A > C
B > C



Search engine:
A > B
C > A
C > B

Evaluation measures

- Simple idea: proportion of correctly ranked pairs
  - Named *ppref*

Assessor ranking Q:
A > B
A > C
B > C

$ppref(Q,R) = \frac{1}{3}$

Search engine R:
A > B
C > A
C > B

Evaluation measures

- They show similar behaviour

	NDCG	*ppref*	*wpref*
DCG	1.00	0.873	0.866
NDCG		0.873	0.866
ppref			0.940

Agreement on system differences.

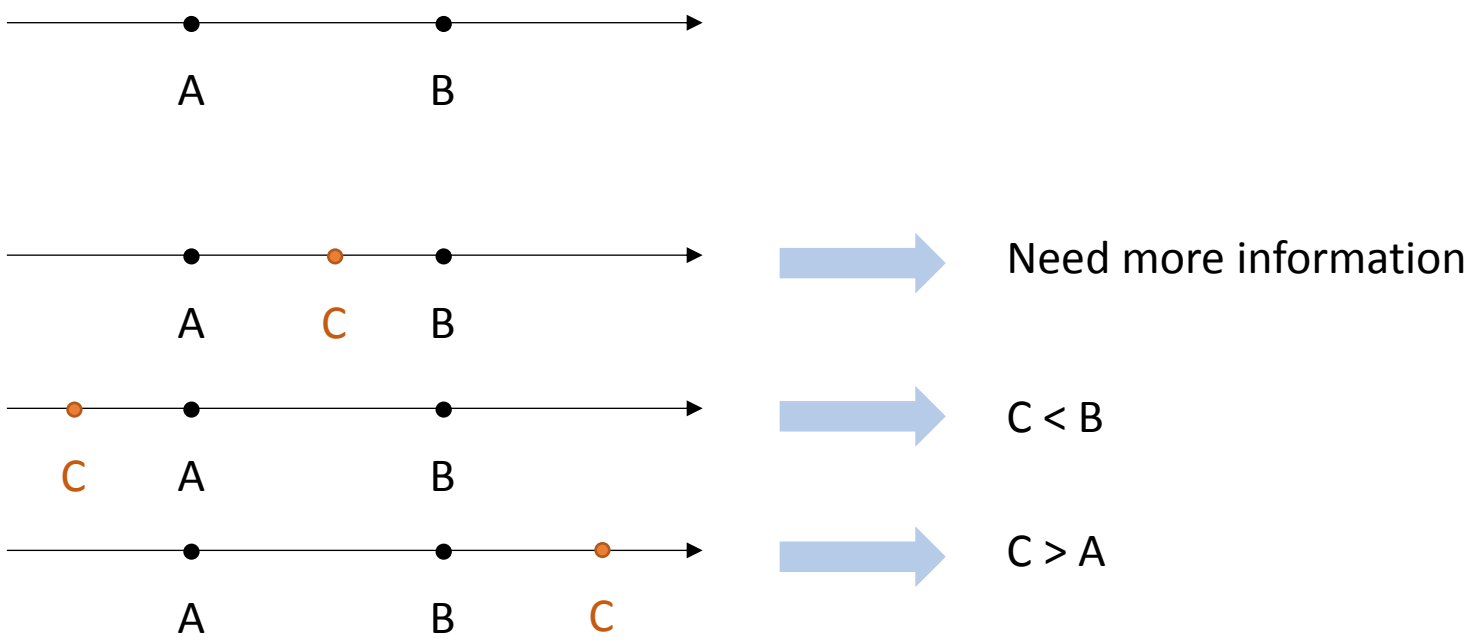
Efficient judging

- Task: Compare two search engines E_1 and E_2
- Problem: $O(n^2)$ pairs for n results

- Solutions:
  - Use transitivity (given > 98 % of the time)
  - Eliminate „bad“ judgments
  - Evaluate pairs (i, j) with high utility first

Efficient judging

- Estimating utility:
  - If A > B in both rankings, value for *ppref* is the same
  - Make use of transitivity:



Efficient judging

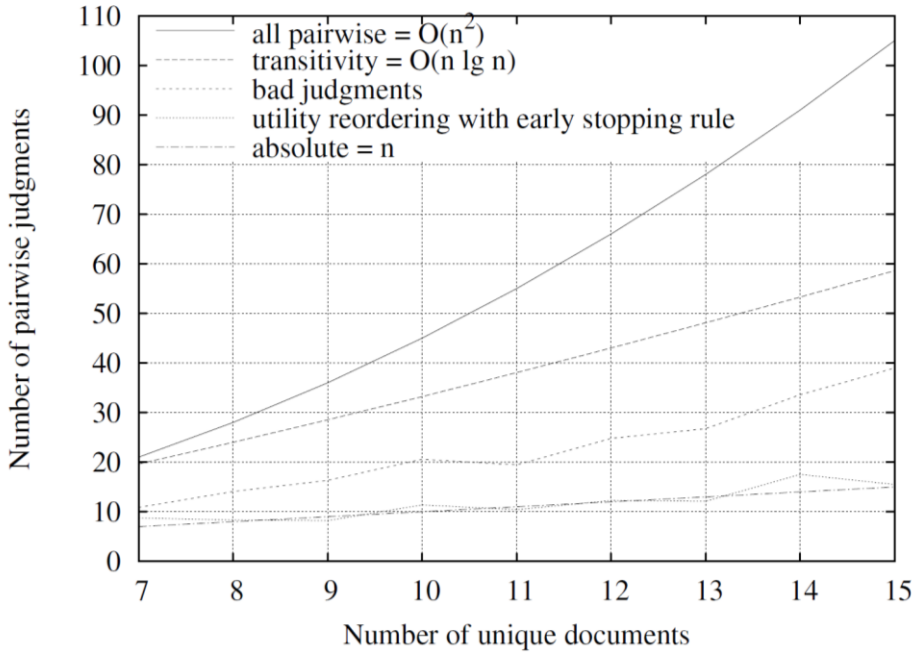
- Gain of a preference pair:
  - Let $R(A > B)$ the subgain we get by just knowing A > B
  - $$G(A > B) = \mathrm{R(A > B)} + \underbrace{\sum_{B'<B} R(B' < A) + \sum_{A'>A} R(A' > B)}_{\text{Transitivity}}$$

- Expected utility of a pair (A,B):
  - $U(A,B) = p(A > B)G(A > B) + p(B > A)G(B > A)$
  - $p(.)$ is assumed to be uniform

Efficient judging

- Then:
  - Choose pair (A,B) with maximum utility at each step
  - Stop when remaining pairs cannot change overall result

- Performance:



Conclusions

- Preferences
  - Are faster and easier for humans to state
  - Cause lower disagreement rates
  - Graded preferences don‘t add value
  - Suitable algorithms (e.g., RankSVM)

- Issues
  - No guarantees for judging heuristics
  - Full evaluation if we just want to compare two search engines?