

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 | C. \$90,000 |
| B. \$45,000 | D. \$100,000 |

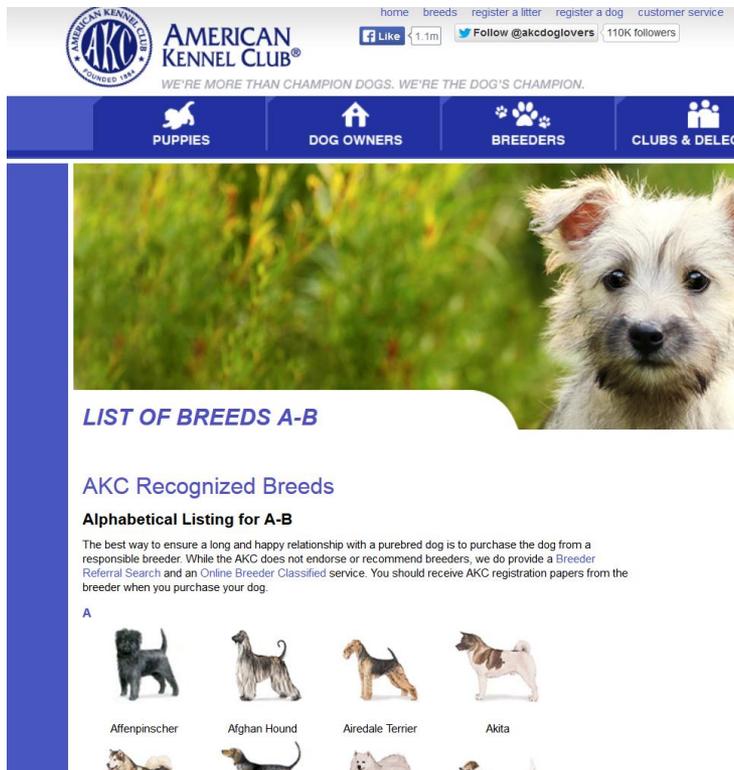
Motivation

- Now: For which car would you pay more?



Motivation

- Query: Dog breeds



The screenshot shows the American Kennel Club (AKC) website. At the top, there is a navigation bar with links for 'home', 'breeds', 'register a litter', 'register a dog', and 'customer service'. Below this is a blue header with icons for 'PUPPIES', 'DOG OWNERS', 'BREEDERS', and 'CLUBS & DELEG'. The main content area features a large image of a white, scruffy dog. Below the image is the heading 'LIST OF BREEDS A-B' and 'AKC Recognized Breeds'. A sub-heading reads 'Alphabetical Listing for A-B'. A paragraph explains that the AKC does not endorse or recommend breeders but provides a Breeder Referral Search and an Online Breeder Classified service. Below this text is a grid of small images of various dog breeds, including an Affenpinscher, Afghan Hound, Airedale Terrier, and Akita.



The screenshot shows a news article from the Houston Chronicle. The top navigation bar includes 'Subscribe to the Houston Chronicle', 'Shopping', 'Classifieds', 'Obits', 'Place an Ad', and 'La Voz'. The date is 'Wednesday March 5, 2014'. The main headline is 'SD Legislature OKs bill against dog breed policies'. Below the headline is the date 'March 4, 2014' and 'Updated: March 4, 2014 3:54pm'. There are social media sharing buttons for 'Comments 0', 'E-mail', 'Print', 'Share 0', 'Tweet 0', 'G+1 0', and 'Pinterest 0'.

PIERRE, S.D. (AP) – The **South Dakota House** has passed a measure to keep local governments from setting policies targeting specific dog breeds.



Four blue buttons are arranged horizontally, labeled 'Fair', 'Good', 'Excellent', and 'Perfect' from left to right.

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

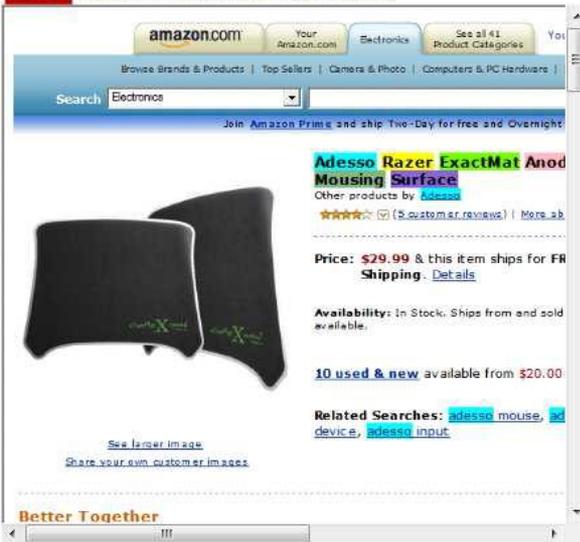
How relevant is this page when you search for:

Adesso Razer ExactMat Anodized Aluminum Mousing Surface (RZ 3050)

End Session | Logout

Fair Good Excellent Perfect

Bad <http://www.amazon.com/.../B0002UECT8>



The screenshot shows an Amazon product page for the 'Adesso Razer ExactMat Anodized Aluminum Mousing Surface (RZ 3050)'. The page includes the Amazon logo, search bar, and product details. The assessor has marked the page as 'Bad' and provided the URL <http://www.amazon.com/.../B0002UECT8>. The product details include a star rating, price of \$29.99, and availability. The page also features a 'Better Together' section at the bottom.

Assessor study

- Graded preferences

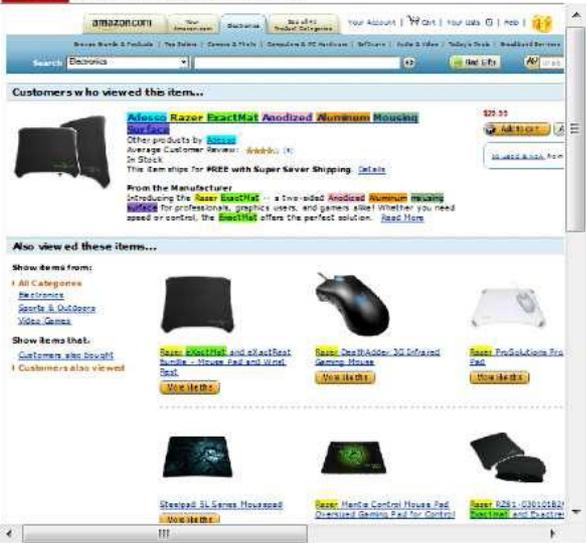
Which page is more relevant when you search for:
Adesso Razer ExactMat Anodized Aluminum Mousing Surface (RZ 3050)

Definitely Here Here Here Definitely Here

Bad <http://www.amazon.com/.../B0002UECT8>



Bad <http://www.amazon.com/.../2>

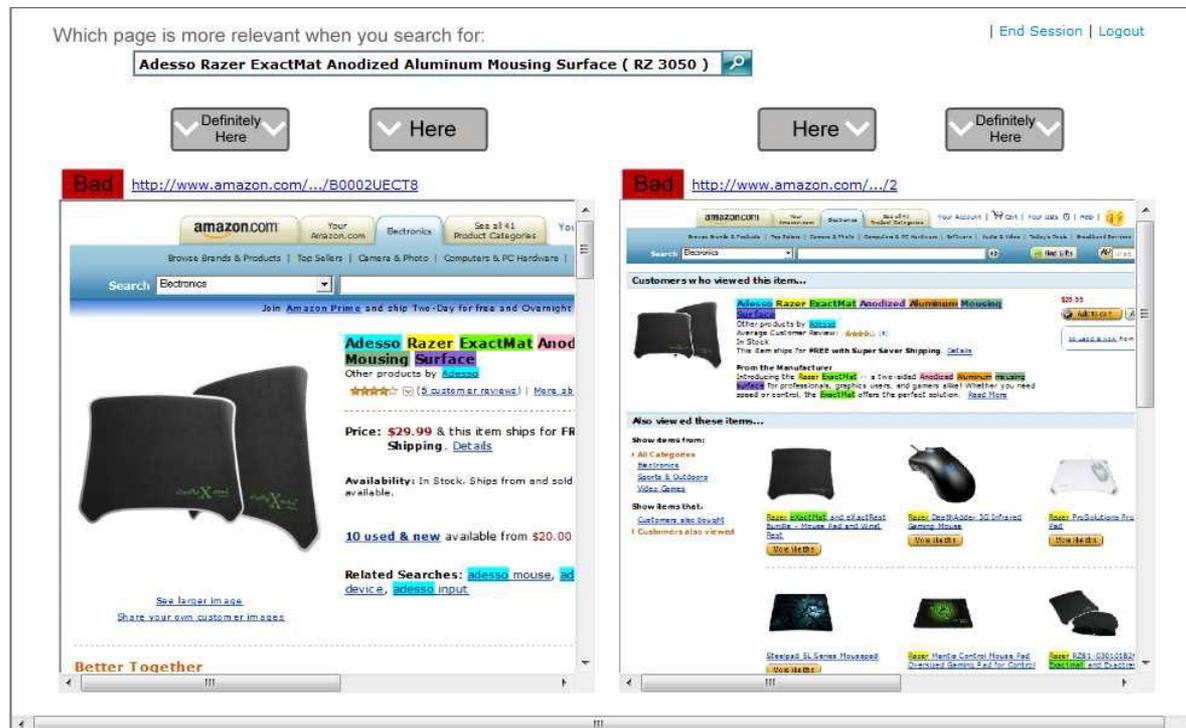


Assessor study

- 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

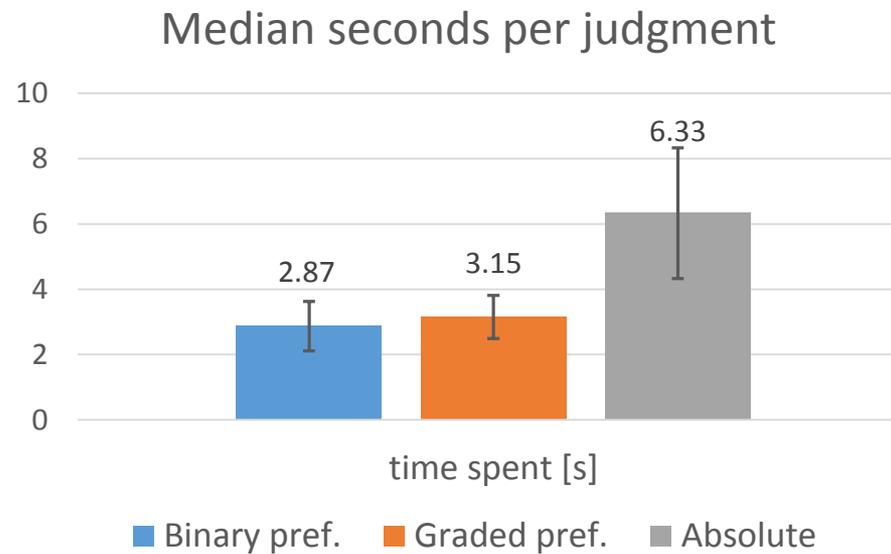
(a) preferences

	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

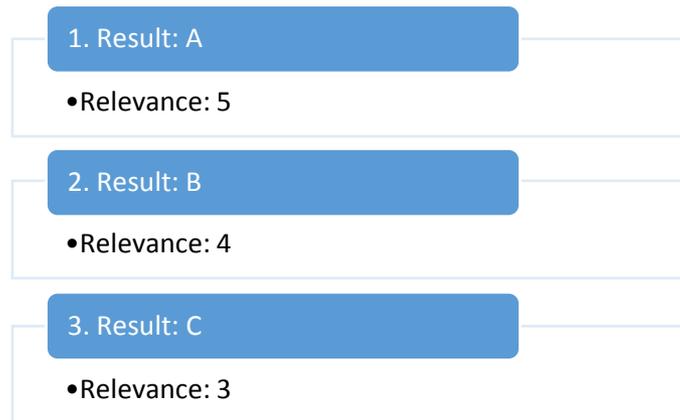
Assessor study

- Making preference statements is faster than making absolute statements

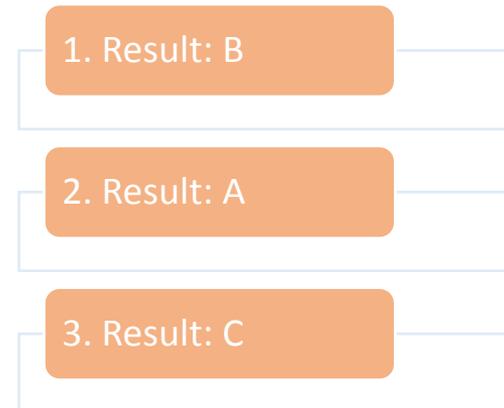


Evaluation measures

- Want to evaluate entire queries
- Traditionally: graded relevance



Gold standard from assessors

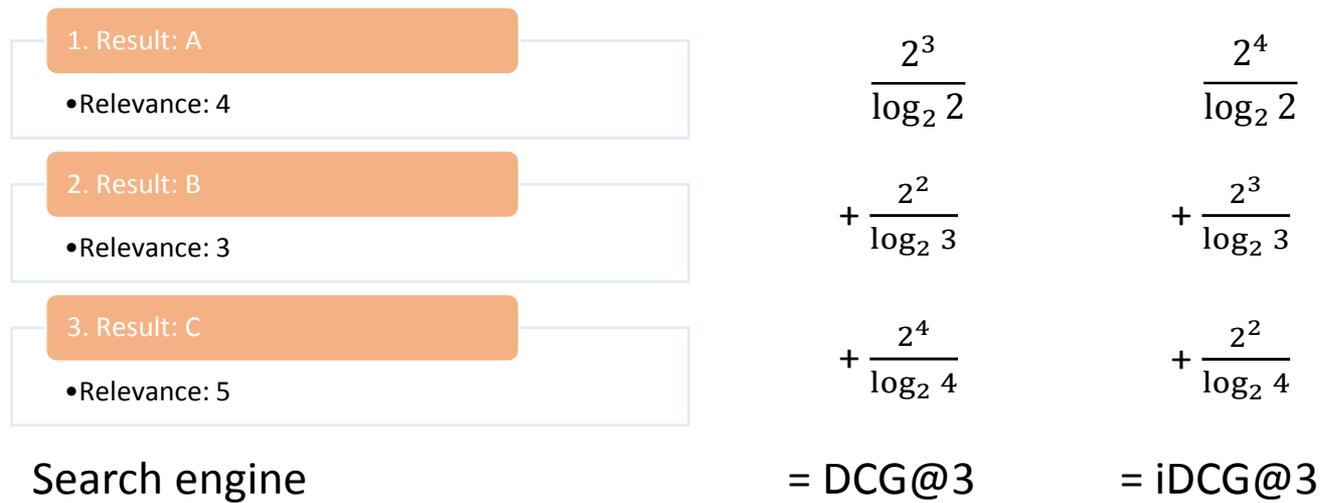


Search engine

Evaluation measures

- DCG (discounted cumulative gain)

- DCG@k = $\sum_{i=1}^k \frac{2^{rel_i-1}}{\log_2 i+1}$
- nDCG@k = $\frac{DCG@k}{iDCG@k}$



Evaluation measures

- Activity: How would you evaluate preferences?

Assessor preferences:

$A > B$

$A > C$

$B > C$

1. Result: A

2. Result: B

3. Result: 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	<i>ppref</i>	<i>wpref</i>
DCG	1.00	0.873	0.866
NDCG		0.873	0.866
<i>ppref</i>			0.940

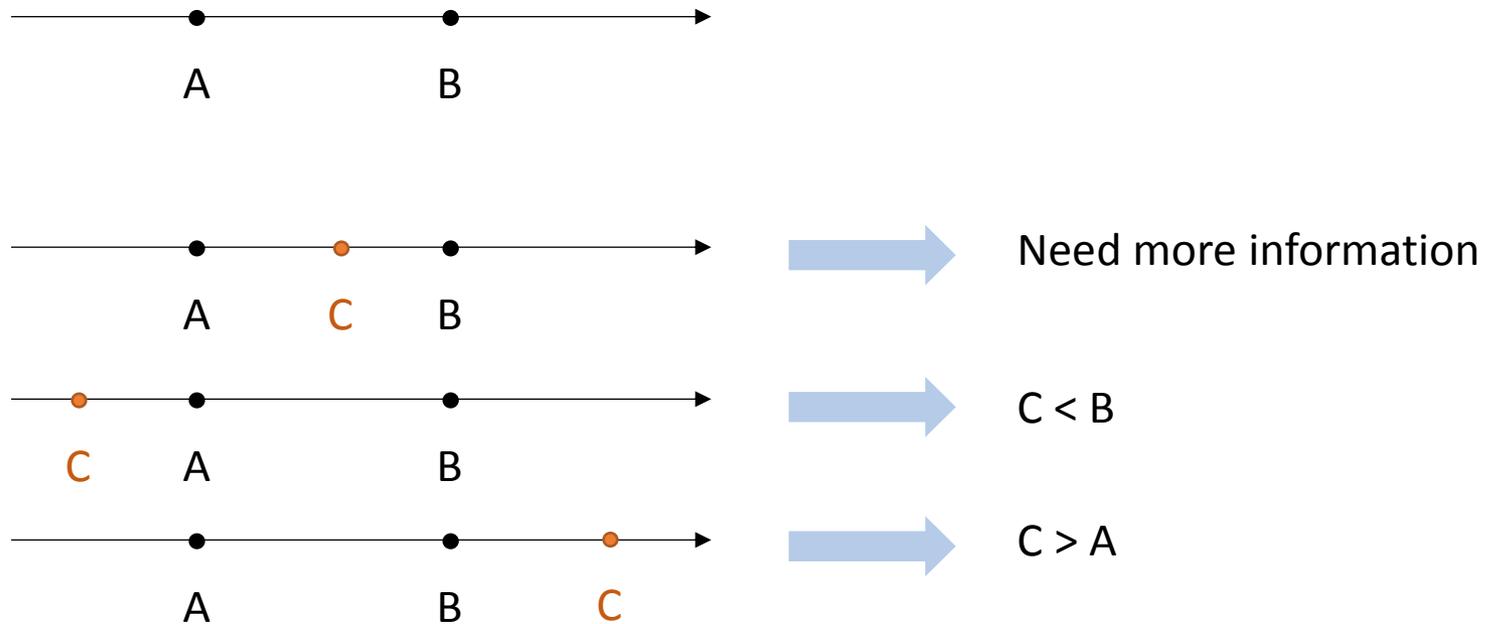
Agreement on system differences.

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 $A > B$

- $$G(A > B) = R(A > B) + \underbrace{\sum_{B' < B} R(B' < A) + \sum_{A' > A} R(A' > B)}_{\text{Transitivity}}$$

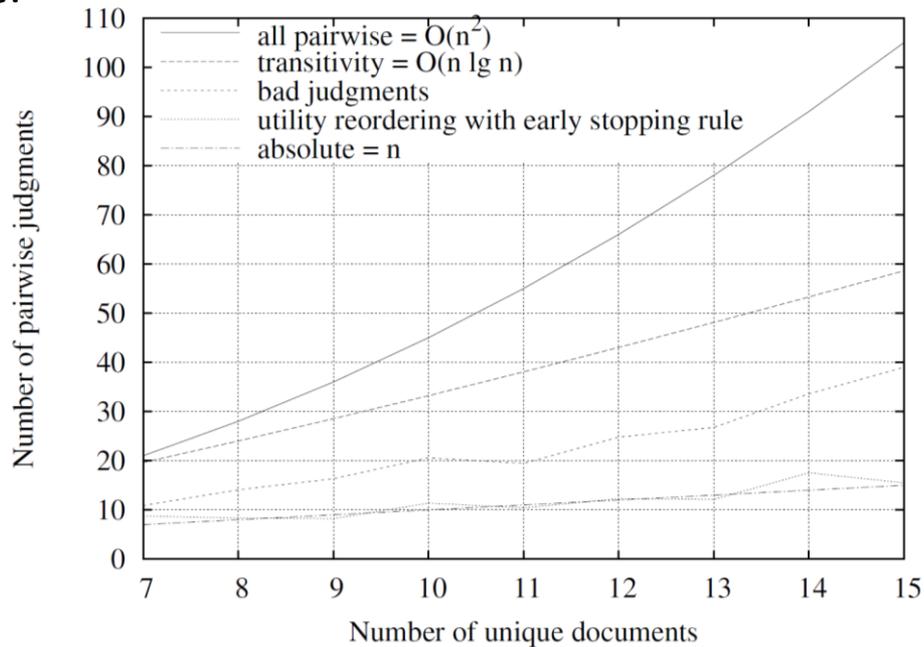


- Expected utility of a pair (A,B):

- $$U(A, B) = p(A > B)G(A > B) + p(B > A)G(B > A)$$
- $p(\cdot)$ 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?