Evaluation II

Mark Sanderson

www.rmit.edu.au

RMIT UNIVERSITY

Content today

- Review exercise you did
- Statistical significance
- Examine test collection design
- Some new evaluation measures
- Building your own test collection
- Other forms of evaluation

Review exercise you did

www.rmit.edu.au

RMIT UNIVERSITY

Results shown in Excel

- Later…

Statistical significance

www.rmit.edu.au

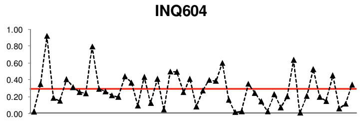
RMIT UNIVERSITY

Take three real systems

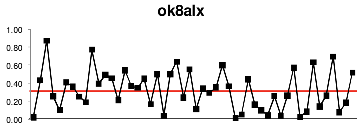
- INQ604
- ok8alx
- CL99XT

- Run them on a TREC test collection
  - Measure MAP
  - INQ604 - 0.281
  - ok8alx - 0.324 (4% absolute; 15% relative)
  - CL99XT - 0.373 (5% absolute; 13% relative)

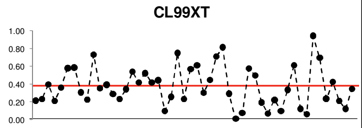
Average?



- INQ604 - 0.281



- ok8alx - 0.324



- CL99XT - 0.373

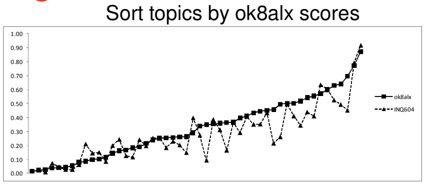
How much is the average

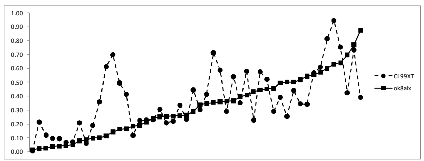
- A product of
  - your IR system
  - chance?
- Slightly different set of topics?
  - Would the average change?

Comparing

- How to decide if difference is worth paying attention to?

Sort topics by ok8alx scores





Comparing two systems

- Significance tests
  - Examine the H_0, null hypothesis
    - "the two systems being compared have effectively the same retrieval characteristics; any difference between them occurred by random chance"
- Most tests consider
  - Mean, standard deviation
  - Compute a probability, p, H_0 that holds.
  - If p is less than a threshold, reject H_0.
    - 0.05 or 0.01

Rejected H_0?

- Some controversy about this, generally
  - Say that H_1 holds
    - Definition depends on 1 or 2 tail test; use 2 tail test
    - H_1: "the two systems being compared are not equal"
    - There is a significant difference between systems

Significance test errors

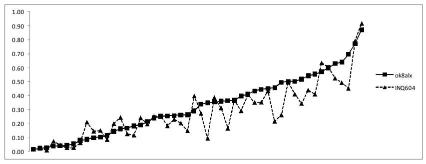
- Type I
  - false positives
- Type II
  - false negatives

Which type of test?

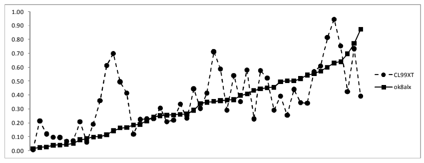
- Non-parametric
  - Sign test
    - Known for type II errors
  - Wilcoxon test
- Parametric
  - T-test
  - Known for type I errors
  - Assumptions of data
  - Get it on Excel
    - (paired test for test collections)

T-test

- p=0.002



- p=0.072



Classic test problems

- Make assumptions about data
  - Often IR data breaks these assumptions
- Evidence tests aren't working well enough
  - Crook, T., Frasca, B., Kohavi, R., & Longbotham, R. (2009). Seven pitfalls to avoid when running controlled experiments on the web. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1105-1114). ACM.
- Good alternatives
  - Don't make assumptions of data
  - Only less used historically
    - Require computational time
  - Bootstrap
    - Sakai, T. (2007). Evaluating Information Retrieval Metrics Based on Bootstrap Hypothesis Tests. *Information and Media Technologies, 2*(4), 1062-1079.
  - Randomisation (permutation) test
    - Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 623-632). ACM New York, NY, USA.

Randomisation test

- Loop for many times {
  - Load topic scores for 2 systems
  - Randomly swap values per topic
  - Compute average for each system
  - Compute difference between averages
  - Add difference to array
  - }
  - Sort array
  - If actual difference outside 95% differences in array
    - Two systems are significantly different

Topic	ok8alx		INQ604
1	0.02		0.02
2	0.43	↔	0.35
3	0.87		0.92
4	0.25	↔	0.18
5	0.10	↔	0.15
6	0.41		0.41
7	0.36	↔	0.31
8	0.25	↔	0.25
9	0.19		0.24
10	0.77	↔	0.79
11	0.40	↔	0.29
12	0.49		0.26
13	0.46		0.21
14	0.21	↔	0.20
15	0.54		0.44
16	0.37		0.36
17	0.35	↔	0.09
18	0.45	↔	0.43
19	0.17		0.12
20	0.50		0.41
21	0.04		0.05
22	0.50	↔	0.50
23	0.64		0.49
24	0.24		0.25
25	0.55	↔	0.41
26	0.11		0.08
27	0.34	↔	0.27
28	0.29		0.40
29	0.35	↔	0.39
30	0.60		0.60
31	0.36	↔	0.16
32	0.01	↔	0.02
33	0.05	↔	0.03
34	0.44		0.35
35	0.16		0.24
36	0.10	↔	0.14
37	0.04		0.03
38	0.26	↔	0.23
39	0.04		0.07
40	0.26	↔	0.20
41	0.57		0.63
42	0.03	↔	0.01
43	0.09		0.21

Mistakes with significance

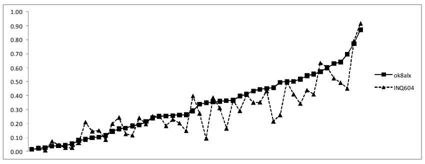
- Do use some form of statistical analysis
  - Significance tests
  - Confidence intervals
  - ANOVA
  - You won't get published if you don't

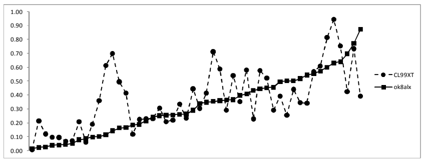
Mistakes with significance

- Are you using a strong baseline?
  - What are you comparing your new system to?

Mistakes with significance

- Is it practically significant?





Mistakes with significance

- Don't search for significance
  - If you keep testing for significance you are guaranteed to find it
  - Remember reject H_0 if ≤5% of it holding
    - 1 in 20 chance it held
  - Run 20 significance tests
    - Very likely one will hold, but will be a false positive
  - Bonferroni correction
    - Shows you how to correctly reduce *p* threshold

Mistakes with significance

	A	B	C	D	E
α	0.25	0.27	0.28	0.15	0.2
β	0.26	0.24	0.28	0.29	0.28
γ	0.21	0.22	0.20	0.24	0.23
δ	0.25	0.23	0.25	0.24	0.22

Mistakes with significance

- Don't use 1 tail tests
  - Unless you know what you are doing
  - Very tempting to use as easier to get significance
    - See Ch 5 of Sanderson 2010 for discussion of tails
- If you fail to get significance
  - Don't give up
  - Use power analysis to understand what the cause might be
    - Webber, W., Moffat, A., & Zobel, J. (2008). Statistical power in retrieval experimentation. *Proceeding of the 17th ACM conference on Information and knowledge management* (pp. 571-580). ACM.

Mistakes with experiments

- When you compare two situations
  - Is there only one thing different?

Examine test collection design

Examine components

- Collection
  - The documents you will search on
- Topics
  - Sampled from query logs or made up
- QRELS
  - Human assessors
- Measure
  - Made up

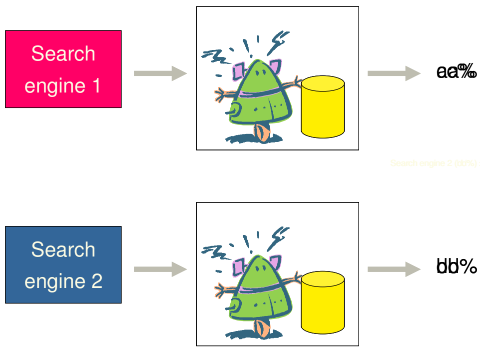
Test collection problems

- Test collection one type of simulation
  - Can newspaper test collection
  - Tell you about web search?
- Possibly not

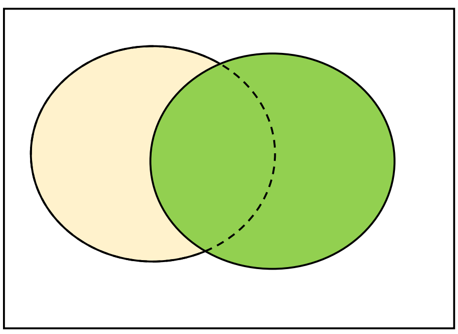
QRELs

- "*A recurring finding from studies involving relevance judgments is that the inter- and intra-judge reliability of relevance judgments is not very high*" - Katter , 1968
- Relevance
  - Personal
  - Changes
    - Over time
    - Over search

Checking QRELS



Examining consistency

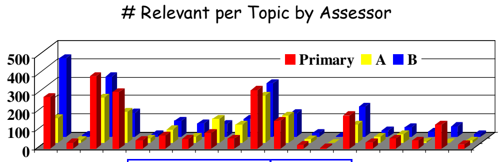


Examining consistency

- Why isn't poor overlap a problem?

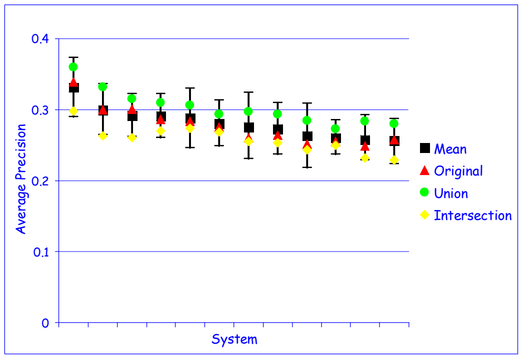
Rank	5	10	20	50	100	250	500	1000
F (%)	82	80	78	75	72	70	69	68

Inter-Judge Agreement



Assessor Group	Overlap
Primary & A	.421
Primary & B	.494
A & B	.426
All 3	.301

Effect of Different Judgments



Checking QRELS

- 1960s
  - Cleverdon
  - Salton

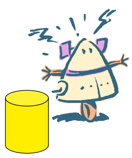
- 1990s
  - Voorhees

- QREL variation doesn't affect relative ranking.

Good enough?

- Continued criticism
  - Really only internal checking
  - Other checks
    - Again internalised checking
  - Ingwersen & Järvelin, 2005
    - Test collection research are working in a cave

Another comparison



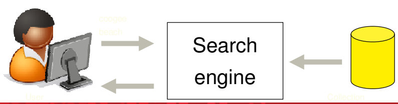
Search engine 2 (bb%) >
Search engine 1 (aa%)

Search engine 2 ( ) >
Search engine 1 ( )

- What do you get the users to do?

Impact of search

- Examine search
  - One search
  - Multiple searches
- Examine broader impact
  - Satisfaction
  - Task that prompted search



User tasks

- Search freely
  - Hersh & Turpin, 2000, 2001, 2002
- Conclusions?
  - Significant difference on test collections?
  - No significant difference in users

More evidence

- Smith & Kantor, 2008
- Two versions of Google
  - Ranks 1-10
  - Ranks 301-310
- Comparisons
  - 301-310: search more
  - Both sets found as many relevant

Conclude from this?

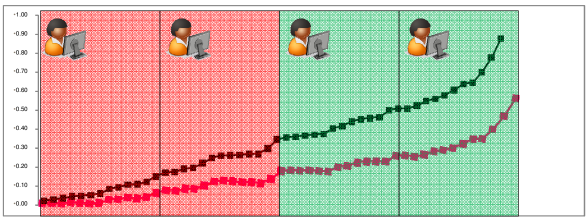
- Searchers are smart
  - Give them a somewhat poorer system they cope

Study users

- Search freely
  - Al-Maskari, Sanderson, SIGIR 2008
- Conclusions?
  - Significant difference on test collections
  - Significant different in users

Why the difference?

- Ensure a clean separation
  - Al-Maskari, Sanderson

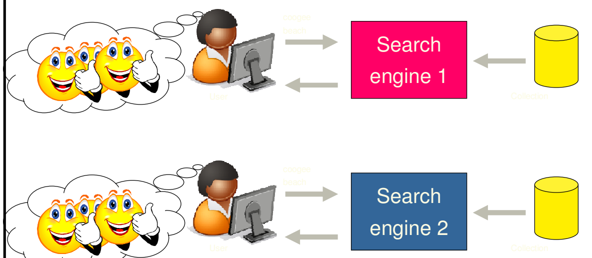


Differences

- Test collections
  - Statistically significant?
- Users?
  - Practically significant?

Joachims – quality bias

Search engine 2 (bb%) > Search engine 1 (aa%)

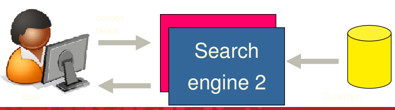


Judging in isolation is hard

- Pepsi Challenge

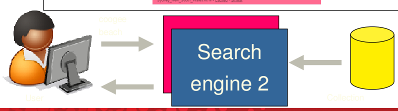
Measure preference

- Side by side



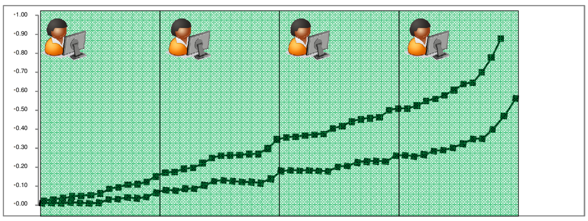
Measure preference

- Interlaced



When comparing

- Can measure more refined user differences



Conclusions on collections

- Test collections appeared to be problematic
- But measurement of people was problematic
  - People can work around flawed IR systems
  - Measure people choices

References

- Al-Maskari, A., Sanderson, M., Clough, P., & Airio, E. (2008). The good and the bad system: does the test collection predict users' effectiveness? *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 59–66). New York, NY, USA: ACM.
- Hersh, W., Turpin, A., Price, S., Chan, B., Kramer, D., Sacherek, L., & Olson, D. (2000). Do batch and user evaluations give the same results? *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 17-24). ACM Press New York, NY, USA.
- Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 154-161). ACM New York, NY, USA.
- Smith, C. L., & Kantor, P. B. (2008). User adaptation: good results from poor systems. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 147-154). ACM New York, NY, USA.
- Thomas, P., & Hawking, D. (2006). Evaluation by comparing result sets in context. *Proceedings of the 15th ACM international conference on Information and knowledge management* (pp. 94-101). ACM Press New York, NY, USA.
- Turpin, A., & Hersh, W. (2001). Why batch and user evaluations do not give the same results. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 225-231). ACM New York, NY, USA.

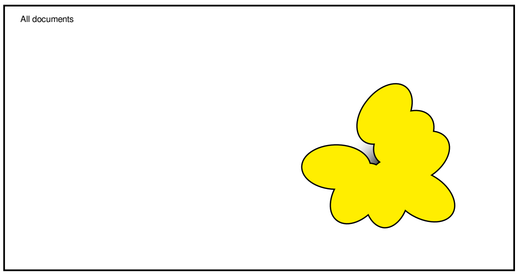
Other evaluation measures

Contents

- Unjudged documents
- Degrees of relevance
- Diversity measures

Unjudged documents

- Pooling



Assume they are not relevant

Can you be cleverer?

$$BPref = \frac{1}{R}\sum_r\left(1 - \frac{|n \text{ ranked higher than } r|}{\min(R,N)}\right)$$

- *R* is the number of documents judged relevant for a particular topic;
- *N* is the number of documents judged not relevant;
- *r* is a relevant retrieved document,
- *n* is a member of the first *R* irrelevant retrieved documents

- Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 25-32). ACM New York, NY, USA.

Bpref – a word of caution

- It was the first such measure
  - Popular
  - Not thought to be that valuable
    - Even by its early proponents
- Alternatives

$$infAP(k) = \frac{1}{R}\sum_r\left[\frac{1}{k} + \frac{(k-1)}{k}\left(\frac{|d100|}{k-1}\bullet\frac{|rel|+\varepsilon}{(|rel|+|nonrel|+2\varepsilon)}\right)\right]$$

- Yilmaz, E., & Aslam, J. A. (2006). Estimating average precision with incomplete and imperfect judgments. *Proceedings of the 15th ACM international conference on Information and knowledge management* (pp. 102-111). ACM Press

Degree of relevance

- Conventional measures
  - Binary relevance only
- Want measure that handle
  - Degrees of relevance
  - Prefer systems that retrieve highly relevant high up the ranking

Discounted Cumulative Gain

– Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS), 20*(4), 422-446.

$$DCG(n) = rel(1) + \sum_{i=2}^{n} \frac{rel(i)}{\log_b(i)}$$

Rank	Rel	Disc	Rel/Disc	DCG
1	2	1.00	2.0	2.0
2	1	1.00	1.0	3.0
3	2	1.58	1.3	4.3
4	0	2.00	0.0	4.3
5	1	2.32	0.4	4.7

Rank	Rel	Disc	Rel/Disc	DCG
1	1	1.00	1.0	1.0
2	0	1.00	0.0	1.0
3	2	1.58	1.3	2.3
4	1	2.00	0.5	2.8
5	2	2.32	0.9	3.6

Alternatives

- Some prefer steeper discount function
  - Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005). Learning to rank using gradient descent. *Proceedings of the 22nd International Conference on Machine Learning*, (pp. 89-96). Bonn, Germany.

$$DCG(n) = \sum_{i=1}^{n} \frac{2^{rel(i)} - 1}{\log(1+i)}$$

nDCG

- DCG not a bounded measure
  - People like [0..1] measures
- Create ideal ranking
  - Normalise DCG against this

$$nDCG(n) = \frac{DCG(n)}{IDCG(n)}$$

Rank	Rel	Disc	Rel/Disc	IDCG
1	2	1.00	2.0	2.0
2	2	1.00	2.0	4.0
3	1	1.58	0.6	4.6
4	1	2.00	0.5	5.1
5	0	2.32	0.0	5.1

Diversity

- Topics can have multiple intents
  - Chicago
    - City tourist information?
    - City weather?
    - Chicago Bears?
    - The band?
- Known about for very long time
  - Verhoeff, J., Goffman, W., & Belzer, J. (1961). Inefficiency of the use of Boolean functions for information retrieval systems. *Communications of the ACM, 4*(12), 557-558.

Recently diverse test collections

- TREC
  - Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., & MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 659-666). ACM New York, NY, USA.
- CLEF
  - Arni, T., Clough, P., Sanderson, M., & Grubinger, M. (2009). Overview of the ImageCLEFphoto 2008 photographic retrieval task. *CLEF'08: Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access* (pp. 500–511). Berlin, Heidelberg: Springer-Verlag.
- Topics cover multiple *nuggets*
  - Also called *aspects* or *sub-topics*

Diversity measure

- α-nDCG, redefines rel function from DCG

$$rel(i) = \sum_{k=1}^{m} J(d_i, k)(1-\alpha)^{r_{k,i-1}}$$

- –m is the number of distinct *nuggets*, $n_1, \ldots, n_m$, relevant to a particular topic;
- –$J(d_i, k)$=1 if an assessor judged that document d_i contained nugget n_k;

$$r_{k,i-1} = \sum_{j=1}^{i-1} J(d_j, k)$$

- –the number of documents ranked before document d_i that were judged to contain nugget n_k

Measures made up?

- All listed evaluation measures never tested on people
  - –Some measure poor for some contexts
    - – Sanderson, M., Paramita, M. L., Clough, P., & Kanoulas, E. (2010). Do user preferences and evaluation measures line up? *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 555–562). New York, NY, USA: ACM. doi:http://doi.acm.org/10.1145/1835449.1835542
- Now exceptions
  - –ERR
    - – Chapelle, O., Metlzer, D., Zhang, Y., & Grinspan, P. (2009). Expected reciprocal rank for graded relevance. *Proceeding of the 18th ACM conference on Information and knowledge management* (pp. 621-630). ACM Press New York, NY, USA.

Building your own test collection

No appropriate test collection?

- Using wrong test collection can produce terrible conclusions
  - –Early TREC web collection
    - –Concluded PageRank worthless!
      - – Hawking, D. (2001). Overview of the TREC-9 Web Track. NIST Special Publication (pp. 87-102). Presented at the Ninth Text Retrieval Conference (TREC-9), Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

What went wrong?

- Test collection
  - –Collection
  - –Topics
  - –QRELS
- Topics
  - –Didn't understand how users search the web
    - –Navigational topics
    - –Informational topics
      - – Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, *36*(2), 3-10. doi:10.1145/792550.792552

Test collection components

- Collection
  - The documents you will search on
- Topics
  - Sampled from query logs
    - Stratified by query type
- QRELS
  - Human assessors
    - Often puts people off

Other ways of finding relevant?

- Web site structure?
  - Harmandas, V., Sanderson, M. and Dunlop, M.D. (1997) Image retrieval by hypertext links, in *the proceedings of the 20th ACM Conference of the Special Interest Group in Information Retrieval (SIGIR)*, 296-303
- Sitemaps
  - Hawking, D. (2004) Challenges in Enterprise Search, in *Proceedings of the Australasian Database Conference* (ADC2004)
- Topic hierarchies
  - Use groupings of documents in Open Directory to locate related documents
    - Haveliwala, T., Gionis, A., Klein, D. and Indyk, P. (2002) Evaluating Strategies for Similarity Search on the Web in *Proc. of the 11th Int. WWW Conference*

More ways?

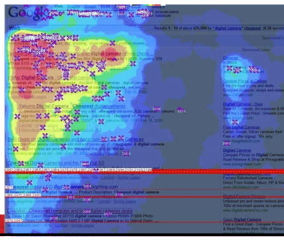
- References?
  - Ritchie, A., Teufel, S., Robertson, S. (2006) Creating a Test Collection for Citation-based IR Experiments, in *Proc of NAACL/HLT conference*
- Temporal clues?
  - Sheridan, Wechsler, Schäuble (1997) Cross-Language Speech Retrieval: Establishing a Baseline Performance, in *Proc. Of ACM SIGIR*

Even more ways?

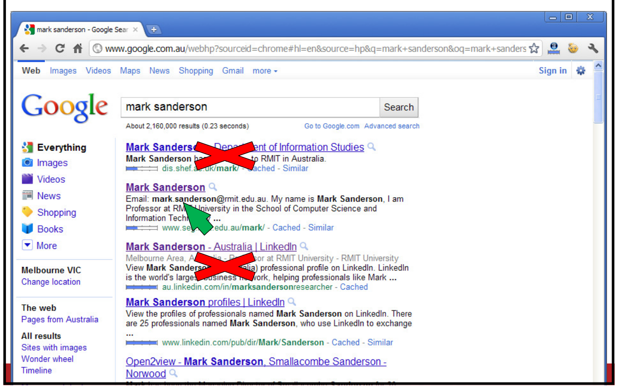
- Display time?
  - Kelly, D., Belkin, N.J. (2004) Display Time as Implicit Feedback: Understanding Task Effects, in *Proceedings ACM SIGIR*
- Clickthroughs
  - Fox, S., Karnawat, K., Mydland, M., Dumais, S., White, T. (2005) Evaluating implicit measures to improve web search, *ACM Transactions on Information Systems*, Vol. 23, No. 2, 147-168
- Tagging and bookmarks
  - Xu, S., Bao, S., FeiB., Su, Z. and Yu, Y. (2008) Exploring Folksonomy for Personalized Search, in *Proceedings ACM SIGIR*

Click data noisy

- Extract preference information
  - Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 154-161). ACM New York, NY, USA.
  - Use eye tracker to observe how user scan result page

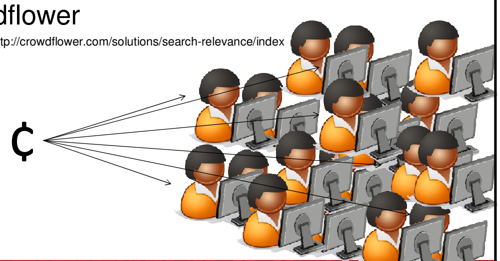


Clues



Other way?

- Amazon's Mechanical Turk?
  - Pay for your QRELs
- Crowdflower
  - http://crowdflower.com/solutions/search-relevance/index



Size of components

- Test theory can help
  - Build a small collection
  - Try your systems out on it
  - Determine accuracy you want to measure
  - Test theory will guide you on size of test collection components
    - Bodoff, D., & Li, P. (2007). Test theory for assessing IR test collections. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 367-374). ACM New York, NY, USA.

Other evaluation

Very brief

Content

- Briefly
  - Other test collections
  - Evaluating using query logs
  - Interactive evaluation
  - Web-based experiments

Learning to rank

- Different needs for evaluation
  - Larger
  - Pre-built feature sets
  - LETOR
    - Liu, T. Y., Xu, J., Qin, T., Xiong, W., & Li, H. (2007). Letor: Benchmark dataset for research on learning to rank for information retrieval. *Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval* (pp. 3-10).
    - Minka, T., & Robertson, S. E. (2008). Selection bias in the LETOR datasets. *SIGIR Workshop on Learning to Rank for Information Retrieval* (pp. 48-51).

Query log studies

- NASA/RECON citation search engine
  - 270,000 citations
    - Meister, D., & Sullivan, D. J. (1967). *Evaluation of user reactions to a prototype on-line information retrieval system* (Prepared under Contract No. NASw-1369 by BUNKER-RAM0 CORPORATION Canoga Park, Calif. No. NASA CR-918). NASA.
  - Examined search logs
    - Number of queries submitted
    - “clicks” on search results
  - Conducted questionnaires with users

Query logs today

- Understand user behaviour
  - Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing and Management, 42*(1), 248-263.
- Evaluating
  - Split log
  - Train on 1st part, test on 2nd
    - Agichtein, E., Brill, E., Dumais, S., & Ragno, R. (2006). Learning user interaction models for predicting web search result preferences. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 3-10). ACM New York, NY, USA.
- Good tutorial on query logs
  - research.microsoft.com/en-us/people/djiang/web_search_and_browse_log_mining.pdf

Interactive evaluation

- Query logs fail to tell you why
  - Conduct smaller experiments where you can understand user behaviour.
- Excellent coverage
  - Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval, 3*(1—2), 1-224.
  - Ingwersen, P., & Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context.* Springer.

Tested your new system

- Can't be sure it will work with user population
  - Need to conduct a test
- A/B testing
  - Show randomly selected group of users
    - New system (B)
    - Old system (A)
    - Monitor logs to look for changes in user behaviour

Excellent further reading

- Microsoft
  - Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery, 18*(1), 140-181.
  - Crook, T., Frasca, B., Kohavi, R., & Longbotham, R. (2009). Seven pitfalls to avoid when running controlled experiments on the web. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1105-1114). ACM.
- Google
  - Tang, D., Agarwal, A., O'Brien, D., Meyer, M. (2010) Overlapping Experiment Infrastructure: More, Better, Faster Experimentation, *Proceedings 16th Conference on Knowledge Discovery and Data Mining*, 2010, pp. 17-26.