

Modeling User Behavior and Interactions

Lecture 2: Interpreting Behavior Data

Eugene Agichtein

Emory University



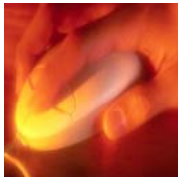
Lecture 2 Plan



- Explicit Feedback in IR
 - Query expansion
 - User control



- From Clicks to Relevance



- 3. Rich Behavior Models
 - + Browsing
 - + Session/Context information
 - + Eye tracking, mouse movements, ...



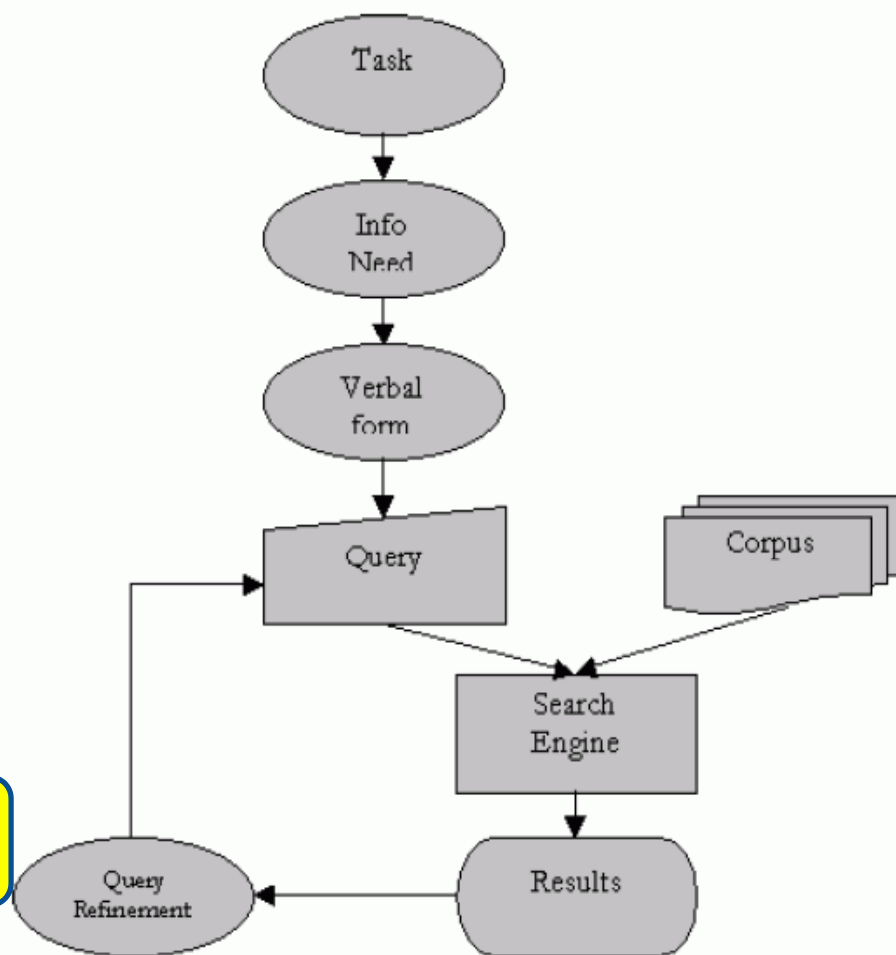


Recap: Information Seeking Process

“Information-seeking ... includes recognizing ... the information problem, establishing a plan of search, conducting the search, evaluating the results, and ... iterating through the process.” - Marchionini, 1989

- Query formulation
- Action (query)
- **Review results**
- **Refine query**

Relevance
Feedback (RF)



Adapted from: M. Hearst, SUI, 2009





Why relevance feedback?

- You may not know what you're looking for, but you'll know when you see it
- Query formulation may be difficult; simplify the problem through iteration
- Facilitate vocabulary and concept discovery
- Boost recall: "find me more documents like this..."





Types of Relevance Feedback

- Explicit feedback: users explicitly mark relevant and irrelevant documents
- Implicit feedback: system attempts to infer user intentions based on observable behavior
- Blind feedback: feedback in absence of any evidence, explicit or otherwise ← will not discuss





Relevance Feedback Example

swine flu russia - Google Search - Mozilla Firefox


File Edit View History Bookmarks Tools Help

http://www.google.com/search?hl=en&client=firefox-a&r ... swine flu

Most Visited Getting Started Latest Headlines ACM Awards

www.russiatoday.com Top Stories & Breaking News Watch TV Feeds Online!

News results for swine flu russia

 **Swine Flu then and now** - 1 hour ago
Wasn't that the **Swine flu** or was it the **Russian flu** and if **Swine flu**, why can't we use the same vaccine we used at that time? Also, what WAS the name of ...
[WSLS.com](#) - 2606 related articles »

Russia contains the H1N1 swine flu virus as it continues to spread ... -
[Telegraph.co.uk](#) - 2534 related articles »

Russia denies entry to visitors with swine flu virus - Russia Now -
[Telegraph.co.uk](#) - 3 related articles »

'Drink whisky to avoid swine flu, Russian fans told - Telegraph
Aug 3, 2009 ... **Russian** football fans have been advised to drink whisky when they go for a football World Cup qualifier next month, ...
[www.telegraph.co.uk/.../swine-flu/.../Drink-whisky-to-avoid-swine-flu-Russian-fans.html](#) - Similar - [Icons]

Russia contains the H1N1 swine flu virus as it continues to spread
Aug 27, 2009 ... According to the World Health Organisation, **Russia** is still "countries not yet hit" by the **swine flu** pandemic.
[www.telegraph.co.uk/.../russiannow/.../Russia-contains-the-H1N1-swine-flu-virus-as-it-continues-to-spread-globally---Russia-Now.html](#) - Similar - [Icons]

Russia takes steps to combat deadly swine flu - RT Top Stories
Apr 26, 2009 ... Meanwhile, **swine flu** cases were confirmed in New York on Sunday as potential cases were reported from New Zealand, Hong Kong, and Spain.
[russiatoday.com/.../Russia_takes_steps_to_combat_deadly_swine_flu_.html](#) - Cached - Similar - [Icons]

Find: fox Next Previous Highlight all Match case

Done

Show options... Results 1 - 25 of 25 similar to [www.telegraph.co.uk/health/swine-flu/5967613](#)

DH home : Department of Health
Official site with collection of publications and policy statements about the National Health Service.
[www.dh.gov.uk/](#) - Cached - Similar

Latest news and features on the NHS and healthcare | Society ...
Aug 28, 2009 ... Ongoing collection of news and features about current issues including diseases and conditions, preventative medicine, the NHS and drug ...
[www.guardian.co.uk/society/health](#) - 6 hours ago - Cached - Similar

Health: Synaesthetes | From the Observer | The Observer
Aug 11, 2002 ... Imagine feeling sounds and hearing colours. Michael Clerizo talks to the synaesthetes about their multi-sensory world.
[www.guardian.co.uk/.../features.magazine97](#) - 21 hours ago - Cached - Similar





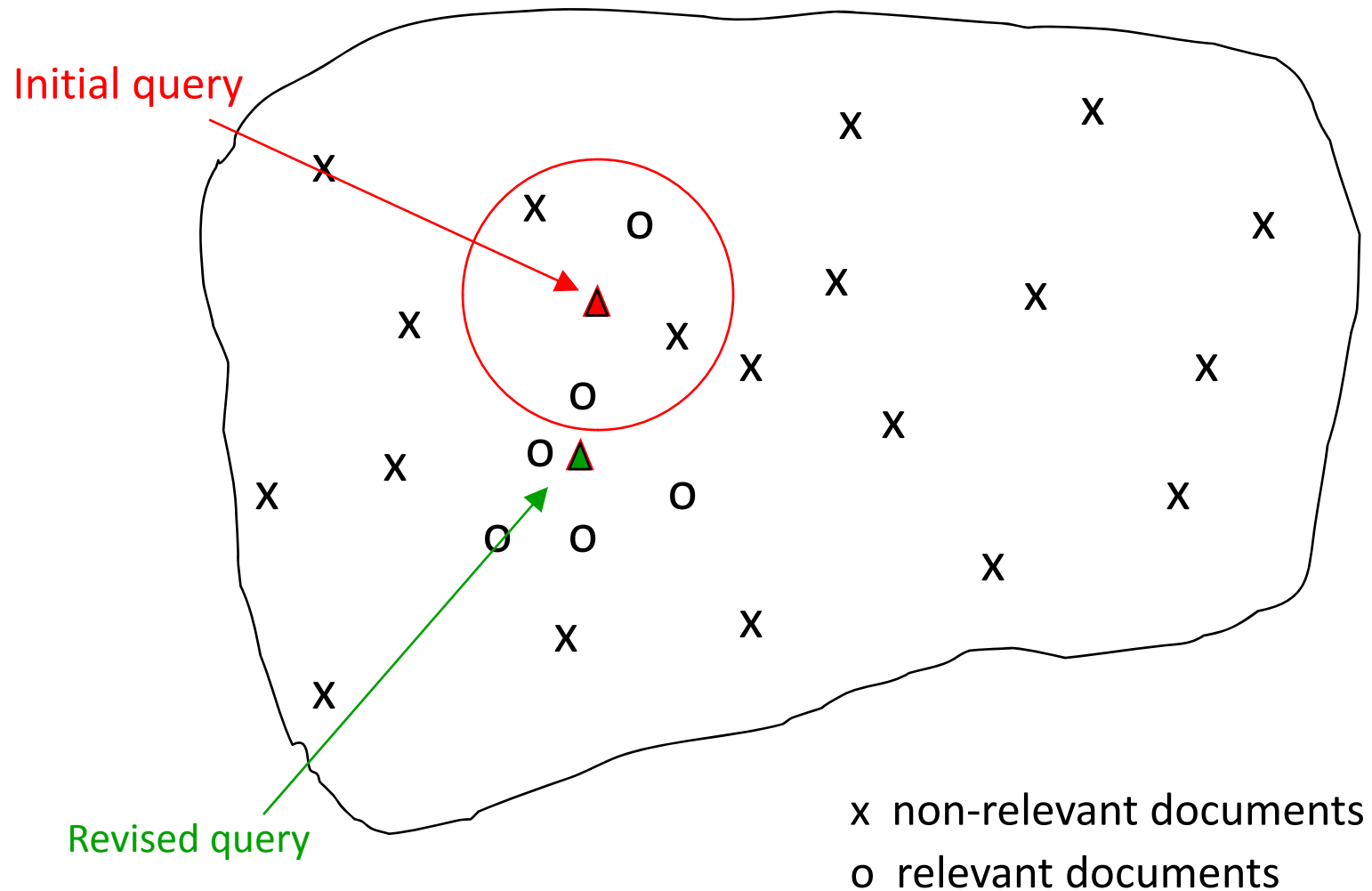
How Relevance Feedback Can be Used

- Assume that there is an optimal query
 - The goal of relevance feedback is to bring the user query closer to the optimal query
- How does relevance feedback actually work?
 - Use relevance information to update query
 - Use query to retrieve new set of documents
- What exactly do we “feed back”?
 - Boost weights of terms from relevant documents
 - Add terms from relevant documents to the query
 - Note that this is hidden from the user





Relevance Feedback in Pictures





Classical Rocchio Algorithm

- Used in practice:

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

q_m = modified query vector;

q_0 = original query vector;

α, β, γ : weights (hand-chosen or set empirically);

D_r = set of known relevant doc vectors;

D_{nr} = set of known irrelevant doc vectors

- New query
 - Moves toward relevant documents
 - Away from irrelevant documents





Rocchio in Pictures

query vector = $\alpha \cdot$ original query vector
+ $\beta \cdot$ positive feedback vector
- $\gamma \cdot$ negative feedback vector

Typically, $\gamma < \beta$

Original query	<table><tr><td>0</td><td>4</td><td>0</td><td>8</td><td>0</td><td>0</td></tr></table>	0	4	0	8	0	0	$\alpha = 1.0$	<table><tr><td>0</td><td>4</td><td>0</td><td>8</td><td>0</td><td>0</td></tr></table>	0	4	0	8	0	0	
0	4	0	8	0	0											
0	4	0	8	0	0											
Positive Feedback	<table><tr><td>2</td><td>4</td><td>8</td><td>0</td><td>0</td><td>2</td></tr></table>	2	4	8	0	0	2	$\beta = 0.5$	<table><tr><td>1</td><td>2</td><td>4</td><td>0</td><td>0</td><td>1</td></tr></table>	1	2	4	0	0	1	(+)
2	4	8	0	0	2											
1	2	4	0	0	1											
Negative feedback	<table><tr><td>8</td><td>0</td><td>4</td><td>4</td><td>0</td><td>16</td></tr></table>	8	0	4	4	0	16	$\gamma = 0.25$	<table><tr><td>2</td><td>0</td><td>1</td><td>1</td><td>0</td><td>4</td></tr></table>	2	0	1	1	0	4	(-)
8	0	4	4	0	16											
2	0	1	1	0	4											
			<hr/>													
			New query	<table><tr><td>-1</td><td>6</td><td>3</td><td>7</td><td>0</td><td>-3</td></tr></table>	-1	6	3	7	0	-3						
-1	6	3	7	0	-3											





Relevance Feedback Example: Initial Query and Top 8 Results

- Query: New space satellite applications

want high recall

- ✓ 1. 0.539, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
- ✓ 2. 0.533, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
- 3. 0.528, 04/04/90, Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
- 4. 0.526, 09/09/91, A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget
- 5. 0.525, 07/24/90, Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
- 6. 0.524, 08/22/90, Report Provides Support for the Critics Of Using Big Satellites to Study Climate
- 7. 0.516, 04/13/87, Arianespace Receives Satellite Launch Pact From Telesat Canada
- ✓ 8. 0.509, 12/02/87, Telecommunications Tale of Two Companies





Relevance Feedback Example: Expanded Query

- | | |
|--------------------|-------------------|
| • 2.074 new | 15.106 space |
| • 30.816 satellite | 5.660 application |
| • 5.991 nasa | 5.196 eos |
| • 4.196 launch | 3.972 aster |
| • 3.516 instrument | 3.446 arianespace |
| • 3.004 bundespost | 2.806 ss |
| • 2.790 rocket | 2.053 scientist |
| • 2.003 broadcast | 1.172 earth |
| • 0.836 oil | 0.646 measure |





Top 8 Results After Relevance Feedback

- ✓ 1. 0.513, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
- ✓ 2. 0.500, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
- 3. 0.493, 08/07/89, When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own
- 4. 0.493, 07/31/89, NASA Uses 'Warm' Superconductors For Fast Circuit
- ✓ 5. 0.492, 12/02/87, Telecommunications Tale of Two Companies
- 6. 0.491, 07/09/91, Soviets May Adapt Parts of SS-20 Missile For Commercial Use
- 7. 0.490, 07/12/88, Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers
- 8. 0.490, 06/14/90, Rescue of Satellite By Space Agency To Cost \$90 Million





Positive vs Negative Feedback

- Positive feedback is more valuable than negative feedback (so, set $\gamma < \beta$; e.g. $\gamma = 0.25$, $\beta = 0.75$).
- Many systems only allow positive feedback ($\gamma=0$).





Relevance Feedback: Assumptions

- A1: User has sufficient knowledge for a reasonable initial query
 - User does not have sufficient initial knowledge
 - Not enough relevant documents for initial query
 - Examples:
 - Misspellings (Brittany Speers)
 - Cross-language information retrieval
 - Vocabulary mismatch (e.g., cosmonaut/astronaut)
- A2: Relevance prototypes are “well-behaved”





A2: Relevance prototypes “well-behaved”

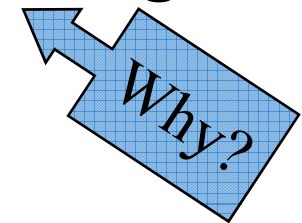
- Relevance feedback assumes that relevance prototypes are “well-behaved”
 - All relevant documents are clustered together
 - Different clusters of relevant documents, but they have significant vocabulary overlap
- Violations of A2:
 - Several (diverse) relevance examples.
 - Pop stars that worked at McDonalds





Relevance Feedback: Problems

- Long queries are inefficient for typical IR engine.
 - Long response times for user.
 - High cost for retrieval system.
 - Partial solution:
 - Only reweight certain prominent terms
Perhaps top 20 by term frequency
- Users are often reluctant to provide explicit feedback
- It's often harder to understand why a particular document was retrieved after relevance feedback





Probabilistic relevance feedback

- Rather than reweighting in a vector space...
- If user marked some relevant and irrelevant documents, then we can build a classifier, such as a Naive Bayes model:
 - $P(t_k | R) = |D_{rk}| / |D_r|$
 - $P(t_k | NR) = (N_k - |D_{rk}|) / (N - |D_r|)$
 - t_k = term in document; D_{rk} = known relevant doc containing t_k ;
 N_k = total number of docs containing t_k
- And then use these new term weights for **re-ranking** the remaining results
- Can also use Language Modeling Techniques (See EDS Lectures)





Empirical Evaluation of RF

- Cannot calculate Precision/Recall on **all** documents

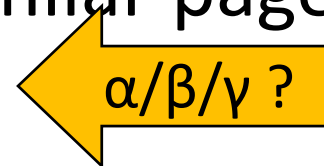


• Must evaluate on documents **not seen by user**

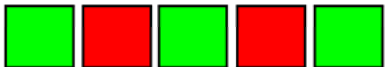
- Use documents in **residual collection** (remove marked docs)
- Final performance often **lower than original query**

- 1 round of relevance feedback is often **very useful**
2 rounds is **sometimes marginally useful**

- Web search engines offer “similar pages” feature:
 - Google (“Similar Documents”)




Review: Common Evaluation Metrics in IR

- **Precision@K**: % relevant in top K results
- Ignores documents ranked lower than K
- Ex: 
 - Prec@3 of 2/3
 - Prec@4 of 2/4
 - Prec@5 of 3/5



Mean Average Precision

- Consider rank position of each relevance doc
 - $K_1, K_2, \dots K_R$
- Compute Precision@K for each $K_1, K_2, \dots K_R$
- Average precision = average of P@K
- Ex:  has AvgPrec of $\frac{1}{3} \cdot \left(\frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right) \approx 0.76$
- MAP is Average Precision across multiple queries



NDCG

- Normalized Discounted Cumulative Gain
- Multiple Levels of Relevance

- DCG:

- contribution of ith rank position: $\frac{2^{y_i} - 1}{\log(i + 1)}$

- Ex:  has DCG score of

$$\frac{1}{\log(2)} + \frac{3}{\log(3)} + \frac{1}{\log(4)} + \frac{0}{\log(5)} + \frac{1}{\log(6)} \approx 5.45$$

- NDCG is normalized DCG
 - best possible ranking as score NDCG = 1





Classical Study: “A Case for Interaction”

Jürgen Koenemann and Nicholas J. Belkin. (1996) A Case For Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness. *CHI 1996*

- Research questions:
 - Does relevance feedback improve results?
 - Is user control over relevance feedback helpful?
 - **Opaque (black box):** User doesn't get to see the relevance feedback process
 - **Transparent:** User shown relevance feedback terms, but isn't allowed to modify query
 - **Penetrable:** User shown relevance feedback terms and is allowed to modify the query
 - How do different levels of user control effect results?





Procedure and Sample Topic

Pretest

Subjects get **tutorial**
on RF

Experiment

Shown 1 mode:

- No RF, opaque, transparent, penetrable
- Evaluation metric used: **precision at 30 documents**

Topic: Tobacco company advertising and the young

Description: A document will provide information on what is a widely held opinion that the tobacco industry aims its advertising at the young.

Narrative: A relevant document must report on tobacco company advertising and its relation to young people. A relevant document can address either side of the question: (1) Do tobacco companies consciously target the young, or (2) As the tobacco industry argues, is this an erroneous public perception. The "young" may be identified as youth, children, adolescents, teenagers, high school students, and college students.





Study Details: Query Interface

Rutgers INQUERY

Reset All UNDO LAST RUN QUERY Show Search Topic Text Show Tutorial Exit RU INQUERY

Enter (next) query term below and hit <RETURN> Clear All Marks You marked 0 documents

Current Query Has 4 term(s):
automobil* manufactur*
car*
defect*
recal*

☐ 1. GM Plans to Recall 62,000 1988-89 Cars With Quad 4 Engines
☐ 2. GM, Ford Recall Vehicles to Repair Defective Parts ---- By Neal Templin S
☐ 3. Isuzu Motors, Honda Commence Car Recalls ---- A Wall Street Journal News I
☐ 4. Ford and GM Recall Series Of Pickup Trucks, Coupes
☐ 5. General Motors Corp. Recalls 196,000 Cars For Defective Brakes

Total of 6747 documents retrieved Jump to rank:

Document # 1 of 6747

GM Plans to Recall
62,000 1988-89 Cars
With Quad 4 Engines

WSJ900413-0013
04/13/90 WALL STREET JOURNAL (I), PAGE B2

DETROIT -- General Motors Corp. said it is recalling 62,000 1988-89 model cars equipped with its high-tech Quad 4 engine to fix defective fuel lines linked to 24 engine fires. GM said the 1988-89 Pontiac Grand Am, Oldsmobile Cutlass Calais and Buick Skylark cars equipped with the 16-valve, four-cylinder Quad 4 engine have fuel lines that could crack or separate from the engines. Although GM has received reports of 24 fires caused by leaks attributable to the faulty fuel lines, a spokesman says the company knows of no injuries resulting from the incidents. GM sold about 312,000 cars equipped with Quad 4 engines in the 1988-89 model years.

In another action, GM said it is recalling about 3,200 of its 1990 Oldsmobile Cutlass Calais and Buick Skylark models to fix fuel-line defects on three engines: the Quad 4, 3.3-liter V-6, and 2.5-liter four cylinder. GM isn't aware of any fires or injuries related to the fuel line problems in this group of cars, the spokesman said.

All repairs will be done free of charge to owners, the company said.

Separately, the U.S. sales arm of Volkswagen AG's Audi subsidiary said it is recalling 1,600 1990-model Audi 80, 90 and Coupe Quattro luxury cars to replace a defective bolt in the assembly that locks the steering when the car is parked. The defective bolt could break, causing the steering wheel to remain locked even after the driver starts the car and begins

Opaque

Penetrable

Reset All UNDO LAST RUN QUERY Show Search Topic

Enter (next) query term below and hit <RETURN> Clear All

Current Query Has 4 term(s):
automobil* manufactur*
car*
defect*
recal*

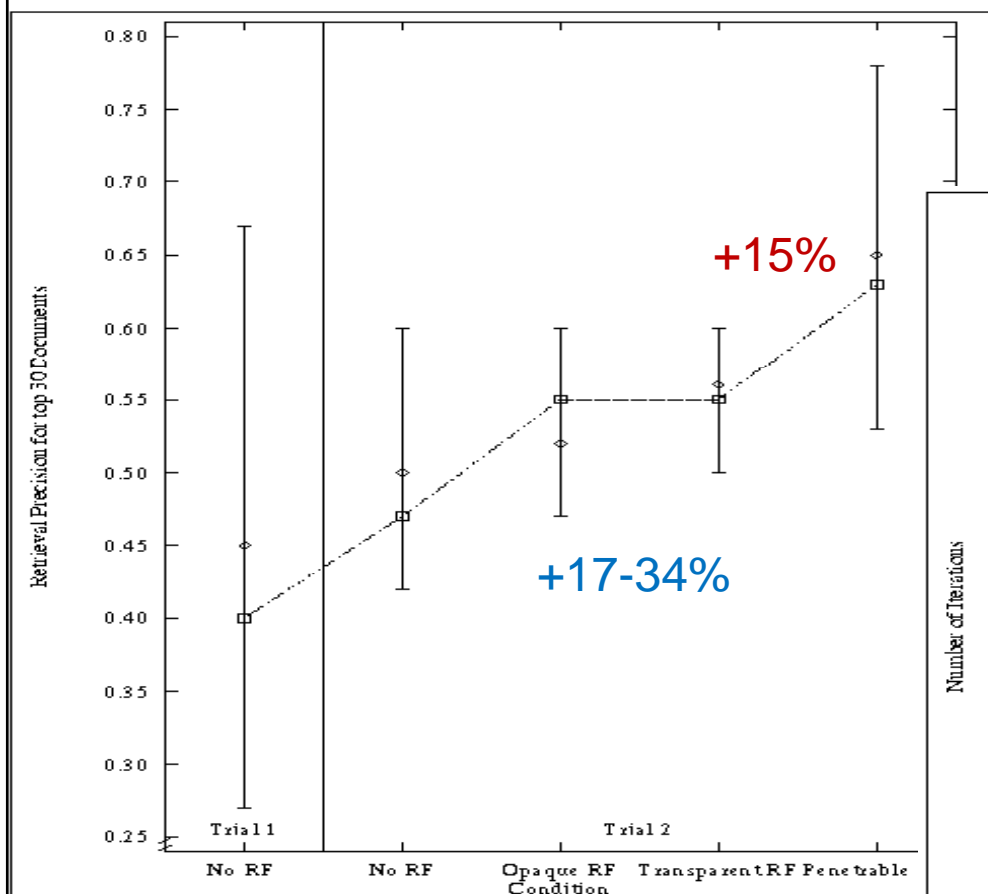
System suggests to add these 9 (stemmed) terms:
accid*
pontiac*
coupe*
fault*
camaro*
cutlass*
leak*
firebird*
oldsmobil*

Penetrable



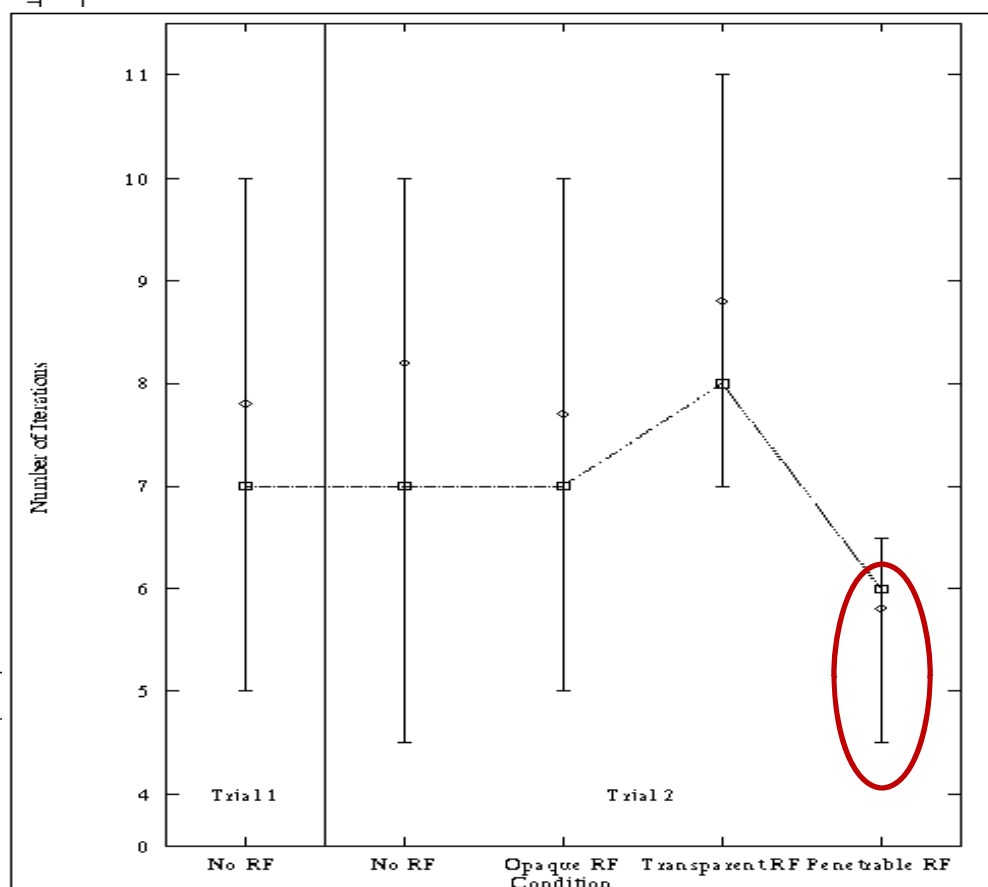


Study Results: Penetrable RF is Best



Penetrable RF performed 15% better than **opaque** and **transparent**

Penetrable interface required **fewer iterations** to arrive at final query





Summary of Explicit Feedback

- Relevance feedback improves results 66% of the time (Spink et al., 2000).
 - Requires ≥ 5 judged documents, otherwise unstable
 - Requires queries for which the set of relevant documents is medium to large
- Only 4% of query sessions used RF “more like this”
 - But, 70%+ stop after first result page, so RF $\sim 1/8$ of rest
- Users more effective at using RF when then can modify expanded query → **Query Suggestion!**



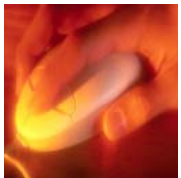
Lecture 2 Plan



- ✓ Explicit Feedback in IR
 - ✓ Query expansion
 - ✓ User control



➤ From Clicks to Relevance



- 3. Rich Behavior Models
 - + Browsing
 - + Session/Context information
 - + Eye tracking, mouse movements, ...





Implicit Feedback

- Users are often reluctant to provide relevance judgments
 - Some searches are precision-oriented (don't need “more like this”)
 - They're lazy or annoyed:
 - “Was this document helpful?”
- Can we gather **relevance** feedback without requiring the user to do anything?
- Goal: **estimate** relevance from behavior





Observable Behavior

Minimum Scope

		Segment	Object	Class
Behavior Category	Examine	View Listen	Select (click)	
	Retain	Print	Bookmark Save Purchase Delete	Subscribe
	Reference	Copy / paste Quote	Forward Reply Link Cite	
	Annotate	Mark up	Rate Publish	Organize





Clicks as Relevance Feedback

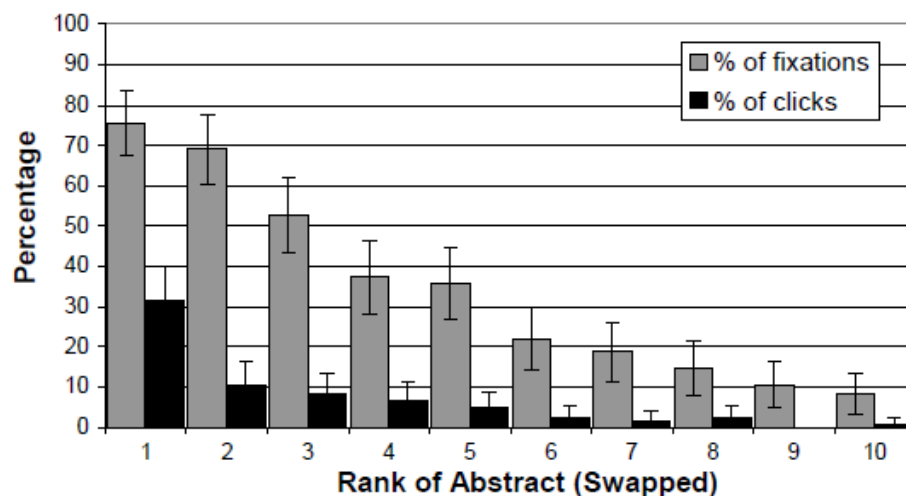
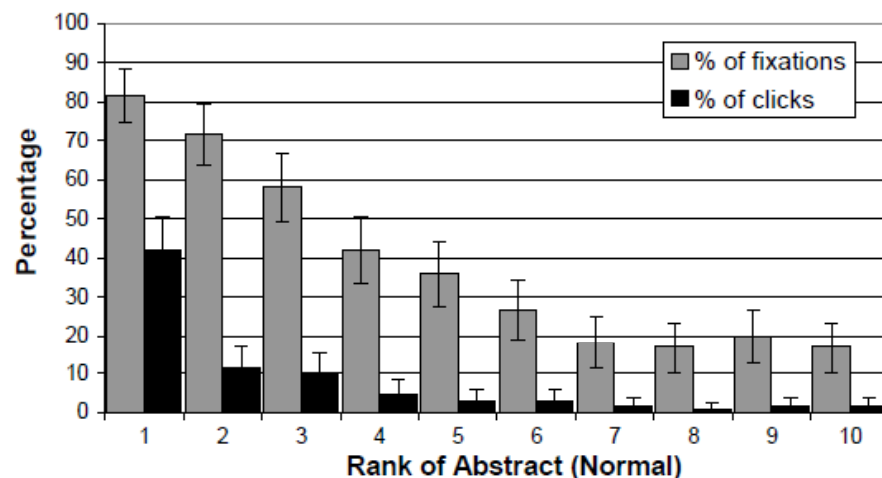
- Limitations:
 - Hard to determine the meaning of a click. If the best result is not displayed, users will click on something
 - Positional bias
 - Click duration may be misleading
 - People leave machines unattended
 - Opening multiple tabs quickly, then reading them all slowly
 - Multitasking
- Compare above to limitations of explicit feedback:
 - Sparse, inconsistent ratings





Interpreting Clickthrough

[Joachims et al., 2005]



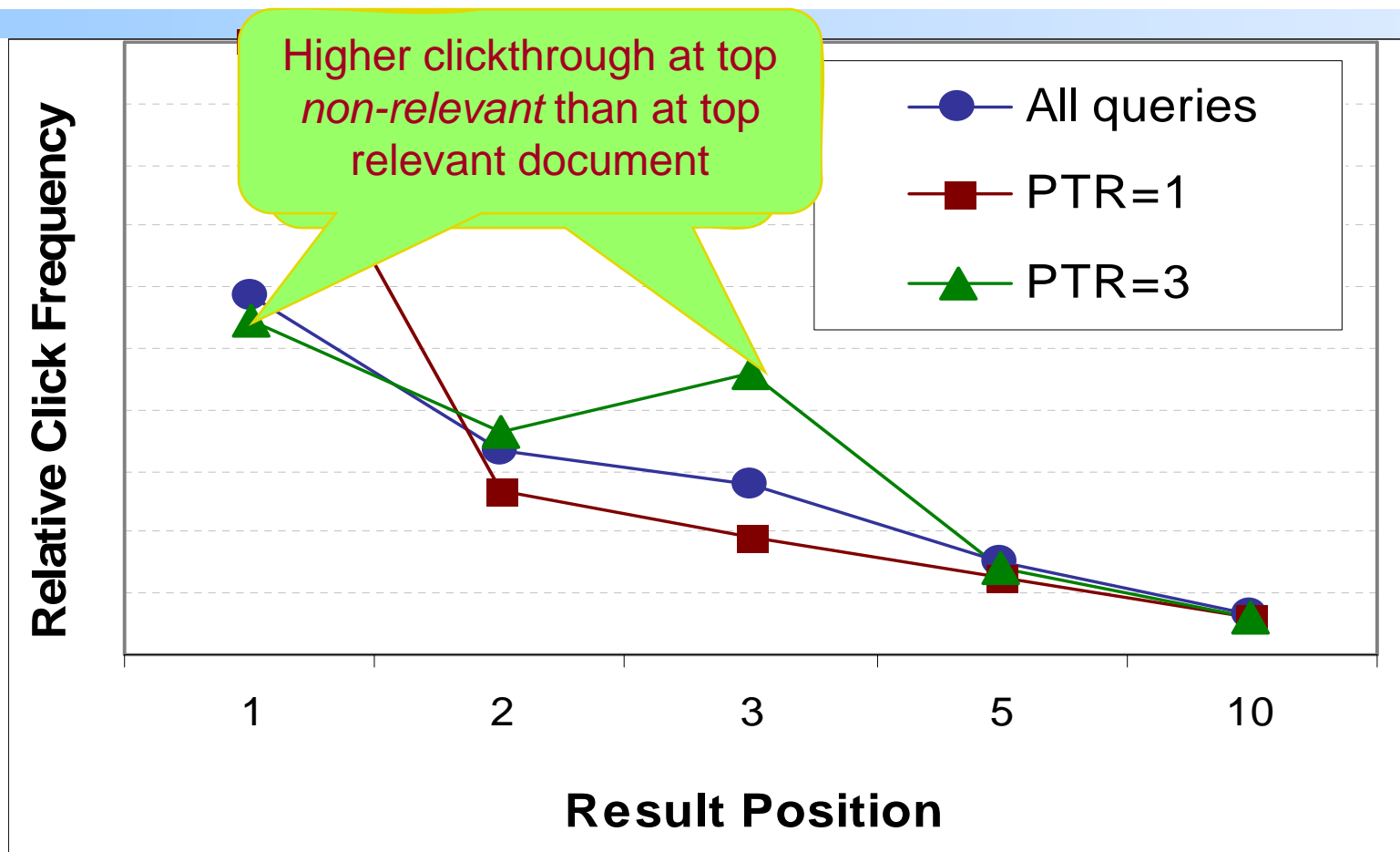
Explicit Feedback Data Strategy	p/q	Abstracts					Pages Phase II all
		Phase I "normal"	"normal"	"swapped"	"reversed"	all	
Inter-Judge Agreem.	N/A	89.5	N/A	N/A	N/A	82.5	86.4
Click > Skip Above	1.37	80.8±3.6	88.0±9.5	79.6±8.9	83.0±6.7	83.1±4.4	78.2±5.6
LastClick > SkipAbove	1.18	83.1±3.8	89.7±9.8	77.9±9.9	84.6±6.9	83.8±4.6	80.9±5.1
Click > Earlier Click	0.20	67.2±12.3	75.0±25.8	36.8±22.9	28.6±27.5	46.9±13.9	64.3±15.4
Click > Skip Previous	0.37	82.3±7.3	88.9±24.1	80.0±18.0	79.5±15.4	81.6±9.5	80.7±9.6
Click > No Click Next	0.68	84.1±4.9	75.6±14.5	66.7±13.1	70.0±15.7	70.4±8.0	67.4±8.2





De-biasing position (first attempt)

[Agichtein et al., 2006]



Relative clickthrough for queries with known relevant results in position 1 and 3

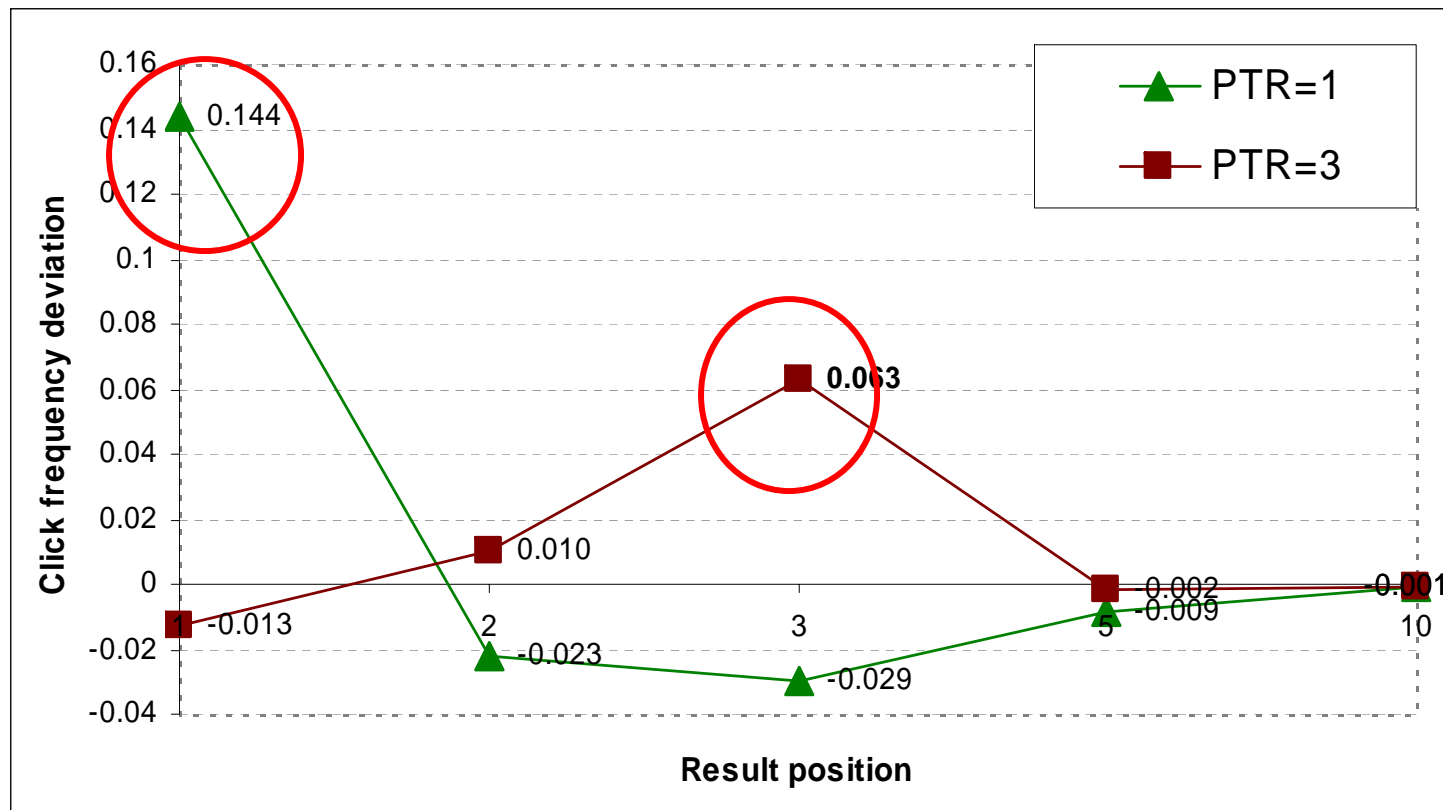




Simple Model: Deviation from Expected

[Agichtein et al., 2006]

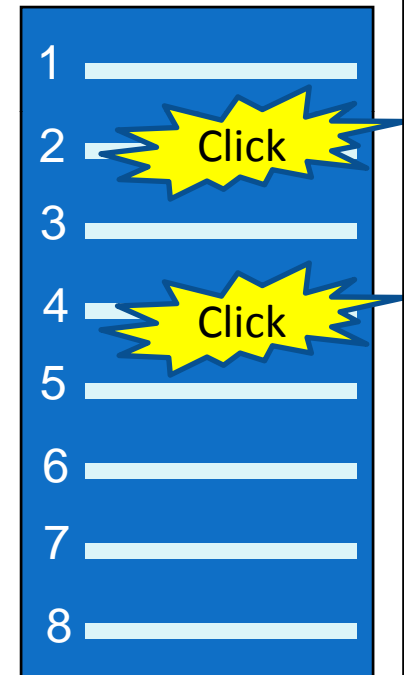
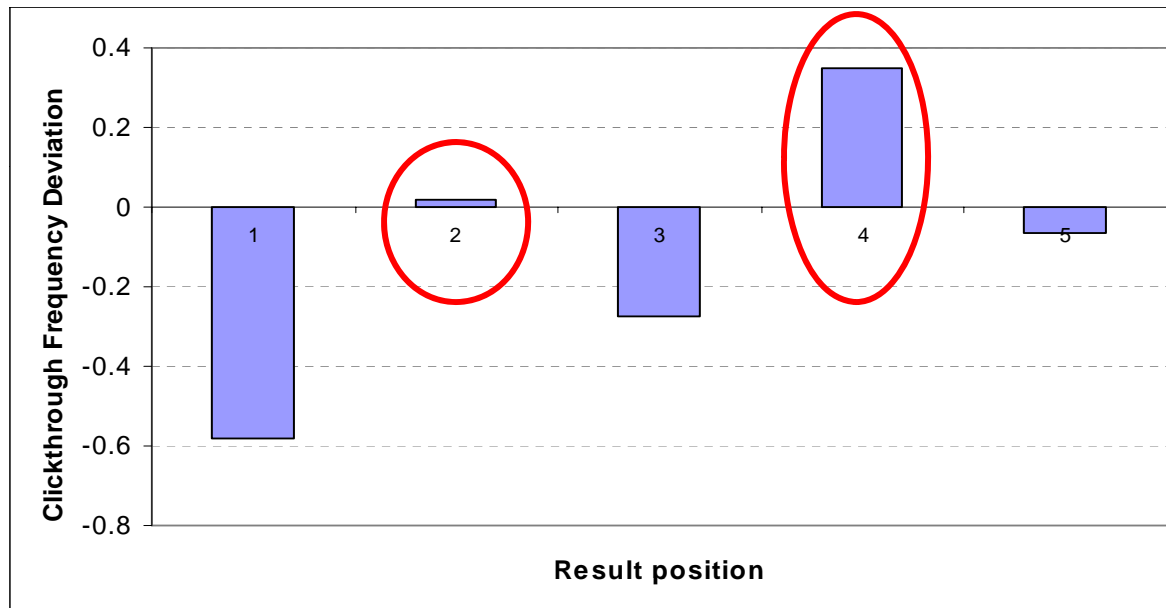
- Relevance component: **deviation** from “expected”:
 $\text{Relevance}(q, d) = \text{observed} - \text{expected}(p)$



Click

Simple Model: Example

- CD: distributional model, extends SA+N
 - Clickthrough considered iff frequency $> \epsilon$ than expected

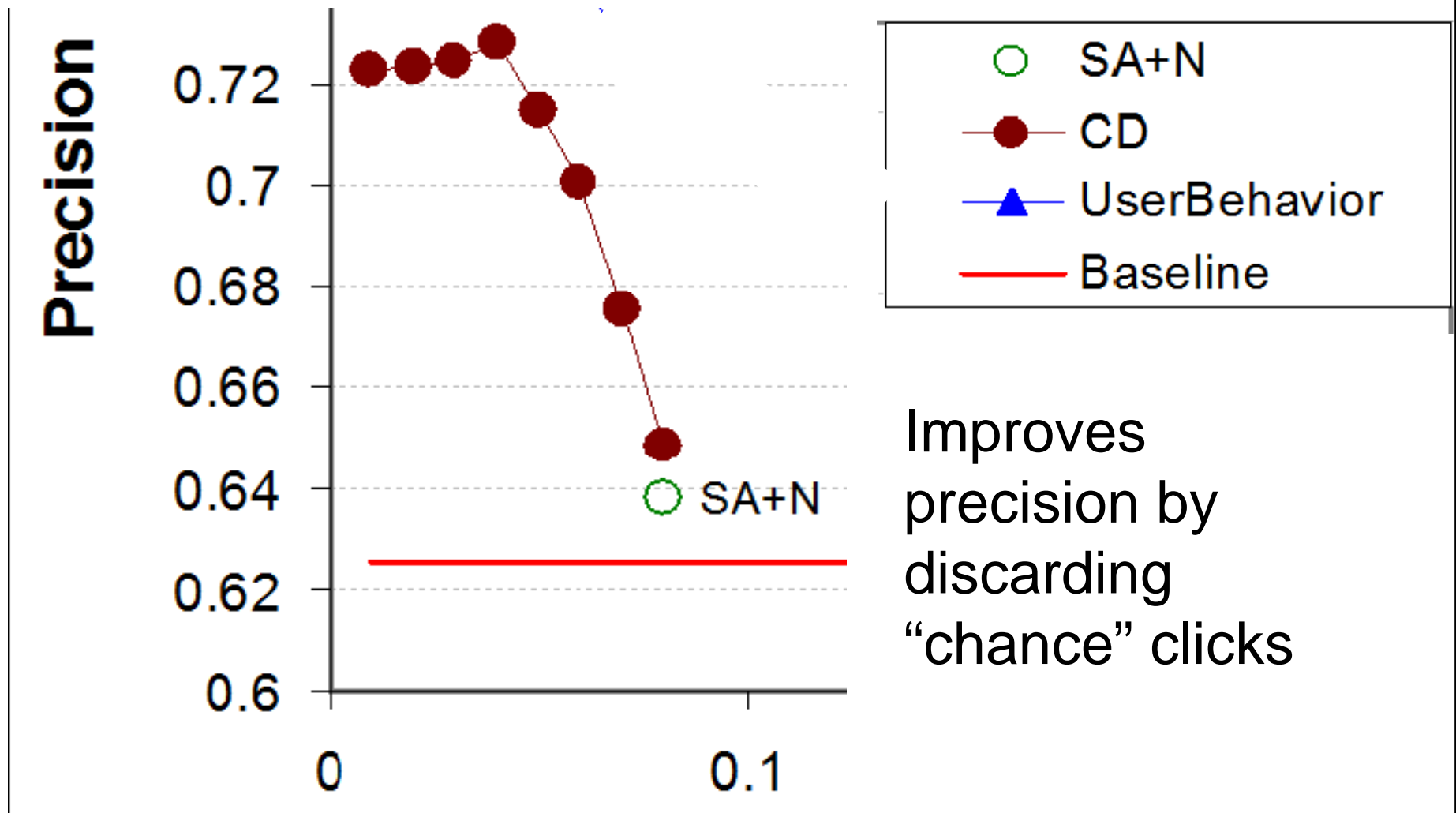


- Click on result 2 likely “by chance”
- $4 > (1, 2, 3, 5)$, but **not** $2 > (1, 3)$





Simple Model Results





Cascade++: Dynamic Bayesian Net

O. Chapelle, & Y Zhang, A Dynamic Bayesian Network Click Model for Web Search Ranking, WWW 2009

did user examine url?

was user satisfied by landing page?

$$A_i = 1, E_i = 1 \Leftrightarrow C_i = 1$$

$$P(A_i = 1) = a_u$$

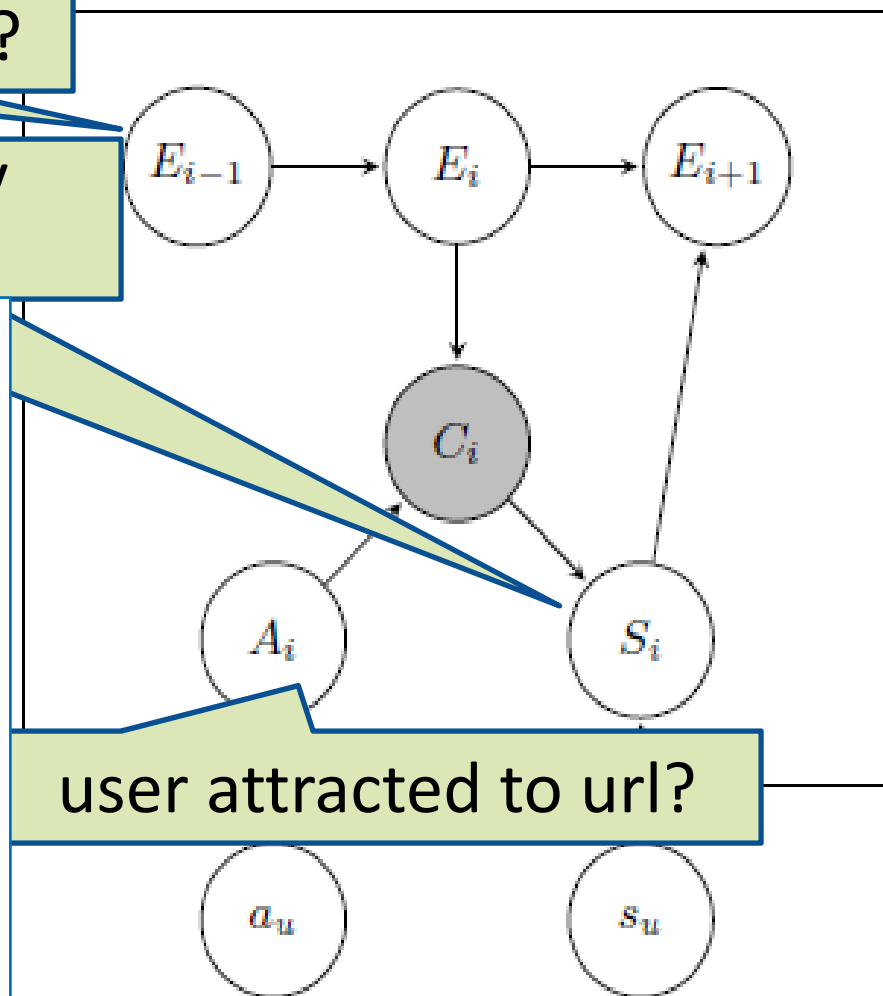
$$P(S_i = 1 | C_i = 1) = s_u$$

$$C_i = 0 \Rightarrow S_i = 0$$

$$S_i = 1 \Rightarrow E_{i+1} = 0$$

$$P(E_{i+1} = 1 | E_i = 1, S_i = 0) = \gamma$$

$$E_i = 0 \Rightarrow E_{i+1} = 0$$





Cascade++: Dynamic Bayesian Net

O. Chapelle, & Y Zhang, A Dynamic Bayesian Network Click Model for Web Search Ranking, WWW 2009

$$A_i = 1, E_i = 1 \Leftrightarrow C_i = 1$$

$$P(A_i = 1) = a_u$$

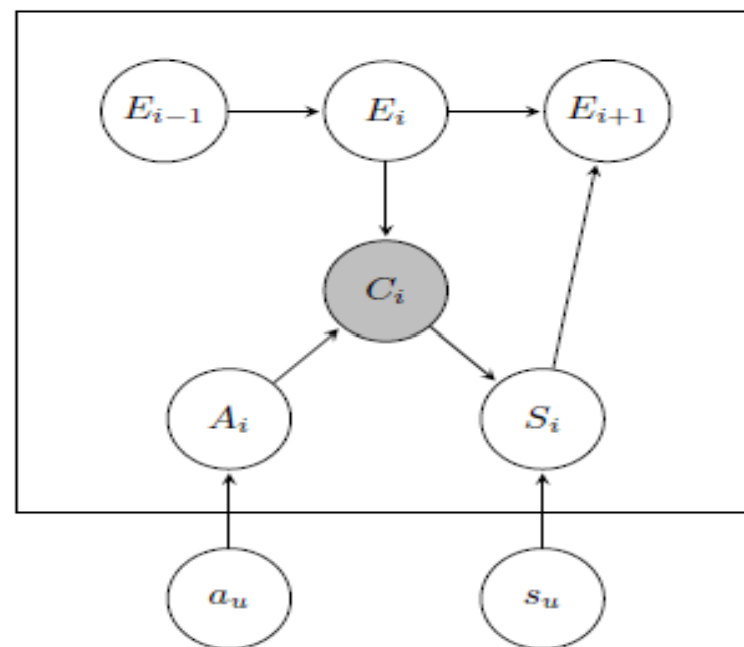
$$P(S_i = 1 | C_i = 1) = s_u$$

$$C_i = 0 \Rightarrow S_i = 0$$

$$S_i = 1 \Rightarrow E_{i+1} = 0$$

$$P(E_{i+1} = 1 | E_i = 1, S_i = 0) = \gamma$$

$$E_i = 0 \Rightarrow E_{i+1} = 0$$



$$r_u := P(S_i = 1 | E_i = 1)$$

$$= P(S_i = 1 | C_i = 1) P(C_i = 1 | E_i = 1)$$

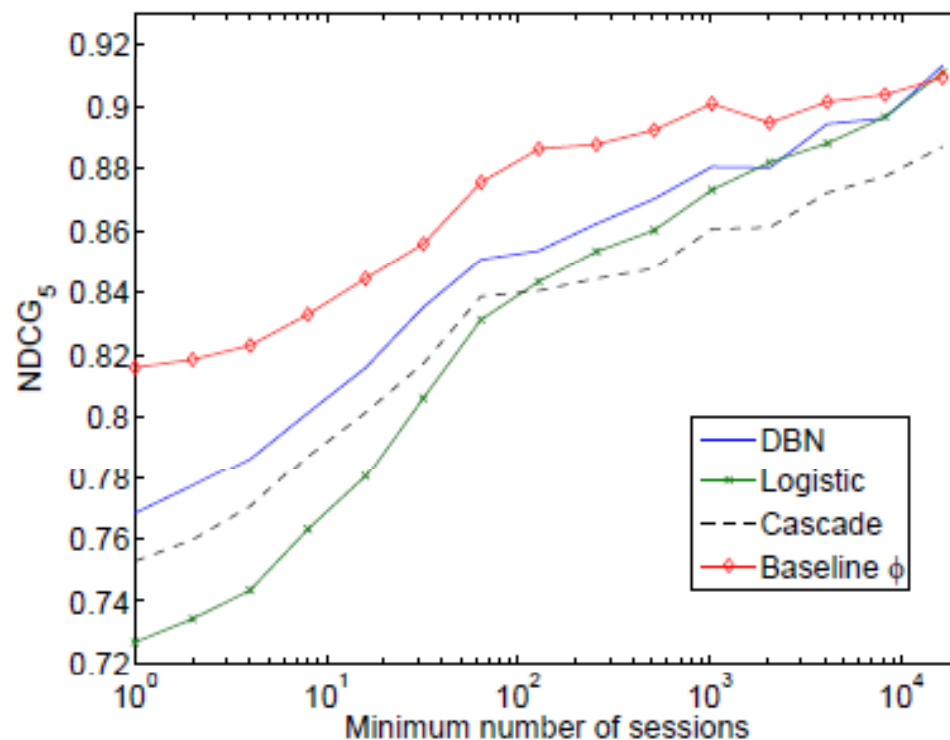
$$= a_u s_u$$



Click Cascade++: Dynamic Bayesian Net (results)

O. Chapelle, & Y Zhang, A Dynamic Bayesian Network Click Model for Web Search Ranking, WWW 2009

Use EM algorithm (similar to forward-backward to learn model parameters; γ set manually



predicted relevance
agrees 80% with
human relevance





Clicks: Summary So Far

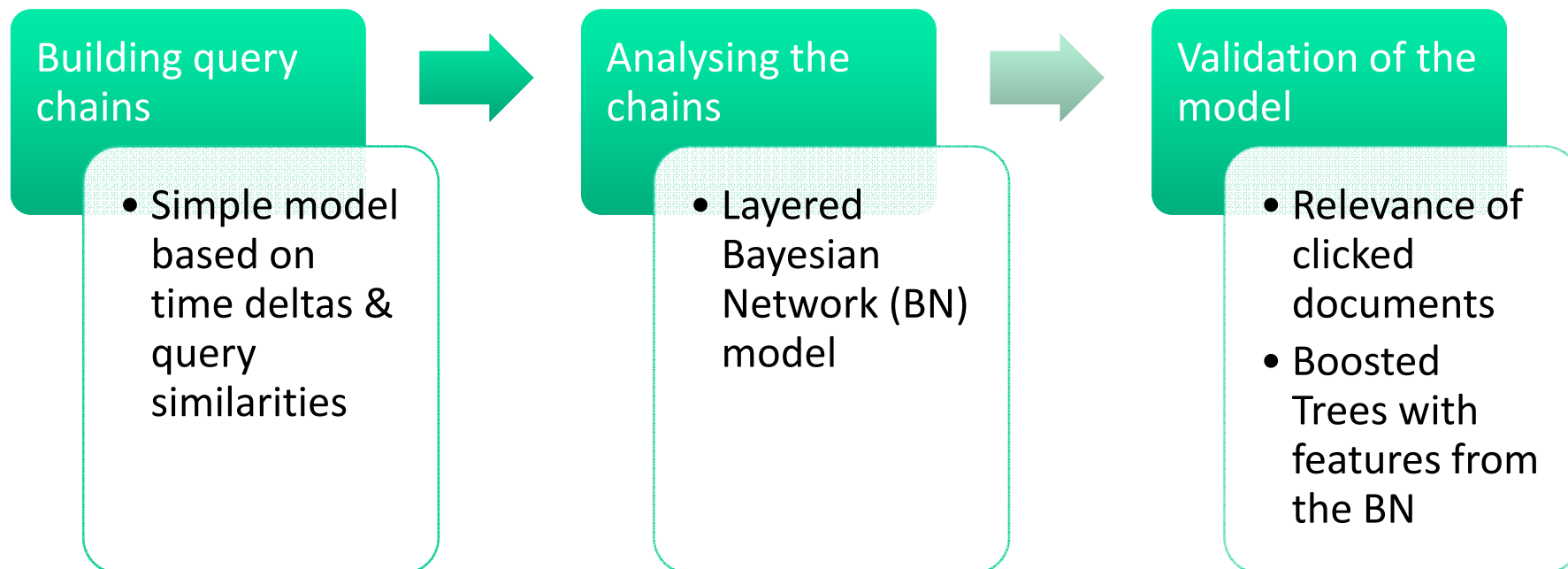
- Simple model accounts for position bias
- Bayes Net model: extension of Cascade model shown to work well in practice
 - Limitations?
- Questions?





Capturing a Click in its Context

[Piwowarski et al., 2009]





Overall process

[Piwowarski et al., 2009]

Grouping atomic sessions

Time threshold

Similarity threshold



(2) world cup · world cup 1998 · quicktime · world cup

(3) world cup · world cup 1998 · quicktime · world cup

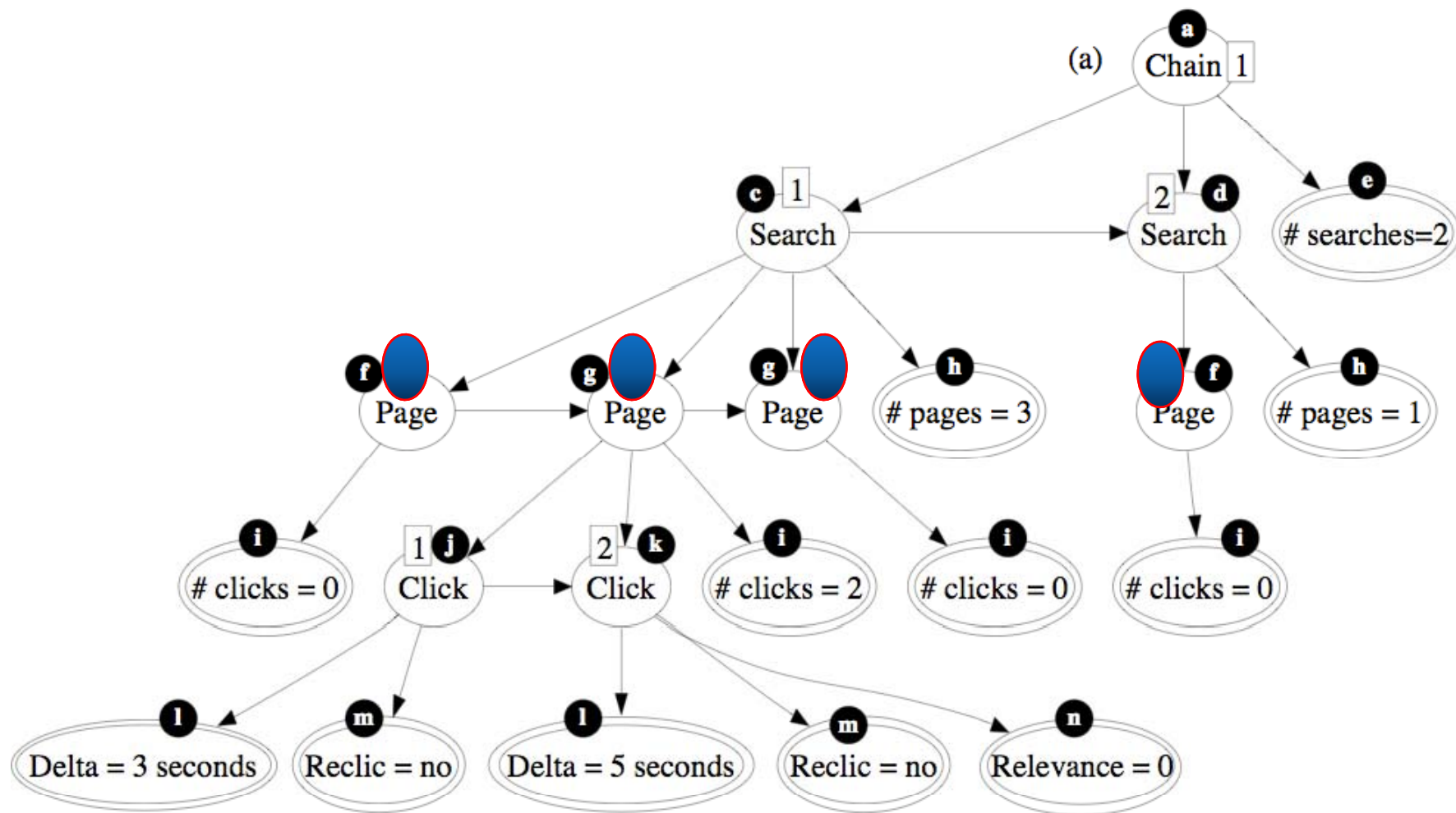
(4) world cup · world cup 1998 · quicktime · world cup





Layered Bayesian Network

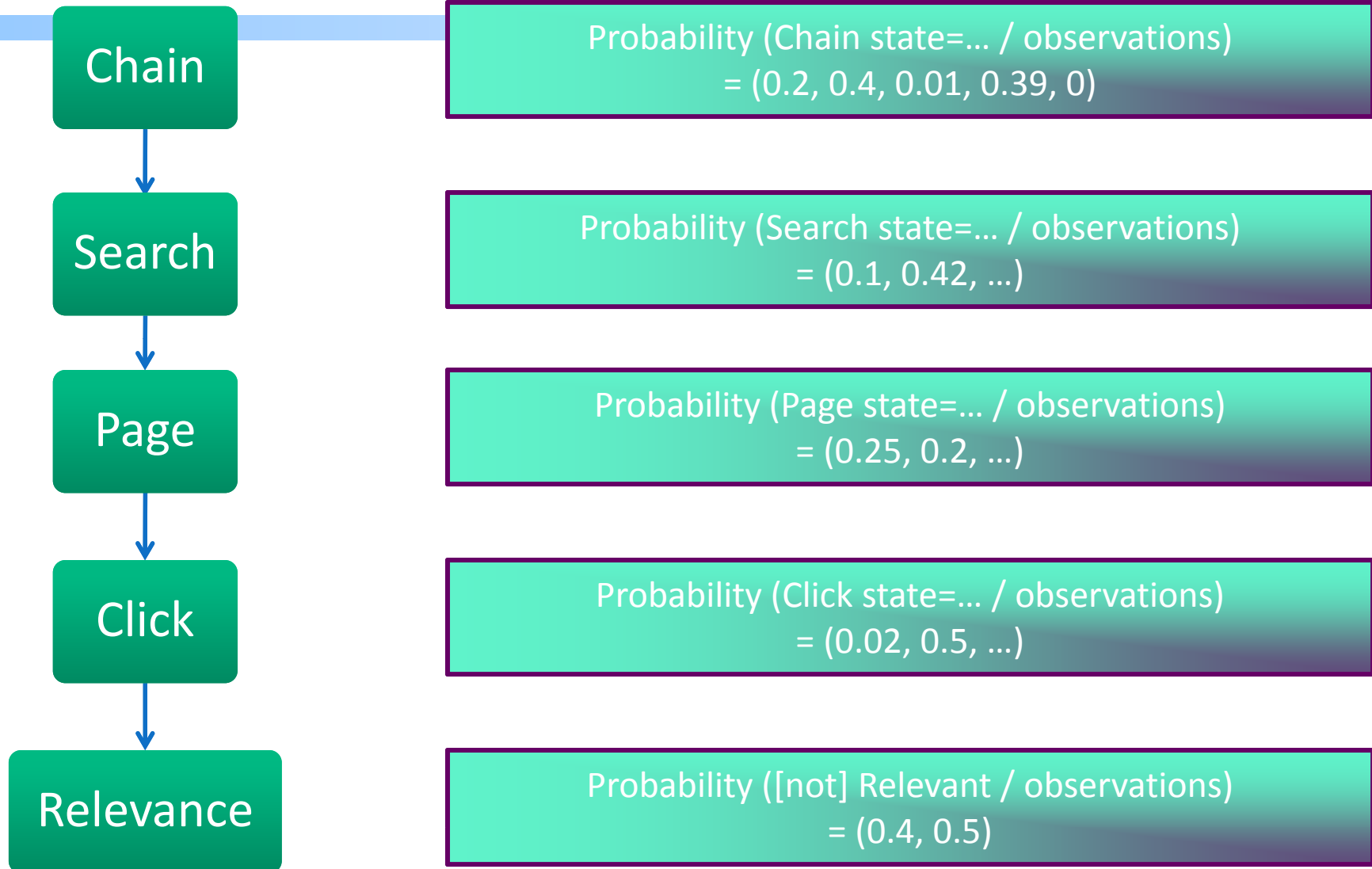
[Piwowarski et al., 2009]





The BN gives the context of a click

[Piwowarski et al., 2009]





Features for one click

[Piwowarski et al., 2009]

- For each clicked document, compute features:
 - (BN) Chain/Page/Action/Relevance state distribution
 - (BN) Maximum likelihood configuration, likelihood
 - Word confidence values (averaged for the query)
 - Time and position related features
- This is associated with a relevance judgment from an editor and used for learning





Learning with Gradient Boosted Trees

[Piwowarski et al., 2009]

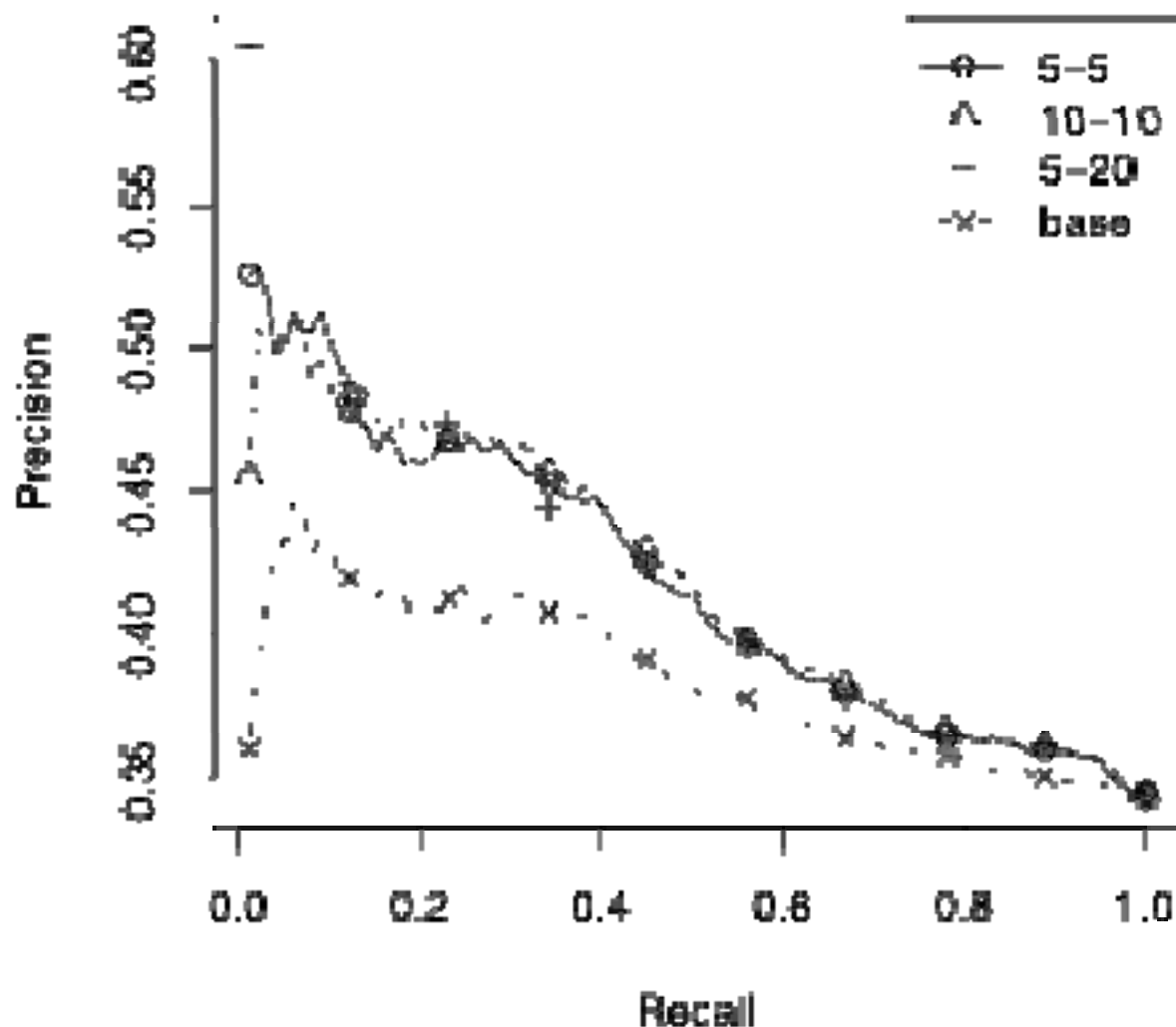
- Use a Gradient boosted trees (Friedman 2001), with a tree depth of 4 (8 for non BN-based model)
- Used disjoint train (BN + GBT training) and test sets
 - Two sets of sessions S1 and S2 (20 million chains) and two set of queries + relevance judgment J1 and J2 (about 1000 queries with behavior data)
 - Process (repeated 4 times):
 - learn the BN parameters on S1+J1,
 - extract the BN features and learn the GBT with S1+J1
 - Extract the BN features and predict relevance assessments of J2 with sessions of S2

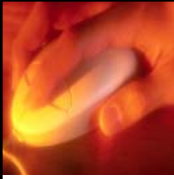




Results: Predicting Relevance of Clicked Docs

[Piwowarski et al., 2009]

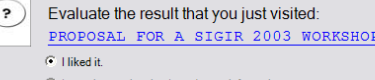




Richer Behavior Models

- Behavior measures of Interest
 - Browsing, scrolling, dwell time
 - How to estimate relevance?
- Heuristics
- Learning-based
 - General model: Curious Browser [Fox et al., TOIS 2005]
 - Query+Browsing model [Agichtein et al., SIGIR 2006]





Evaluate the result that you just visited:

[PROPOSAL FOR A SIGIR 2003 WORKSHOP...](#)

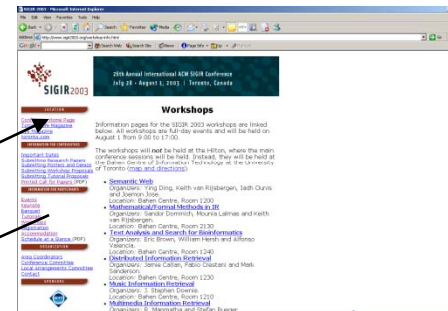
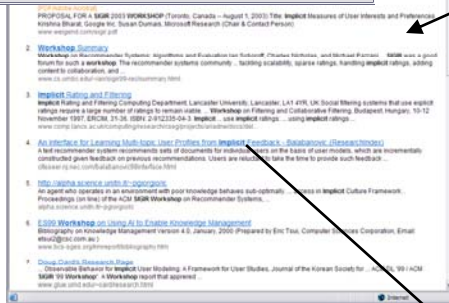
☒ I liked it.

☐ It was interesting, but I need more information.

☐ I didn't like it.

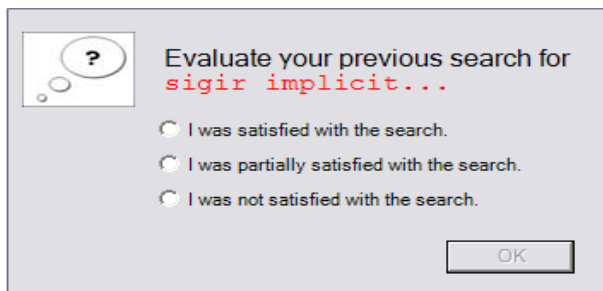
☐ I did not get a chance to evaluate it (broken link, foreign language, etc.).

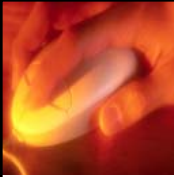
OK



☐ An entirely new search?

☐ A refinement of your previous search?





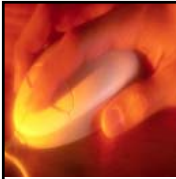
Data Analysis

[Fox et al., 2003]

- Bayesian modeling at result and session level
- Trained on 80% and tested on 20%
- Three levels of SAT – VSAT, PSAT & DSAT
- Implicit measures:

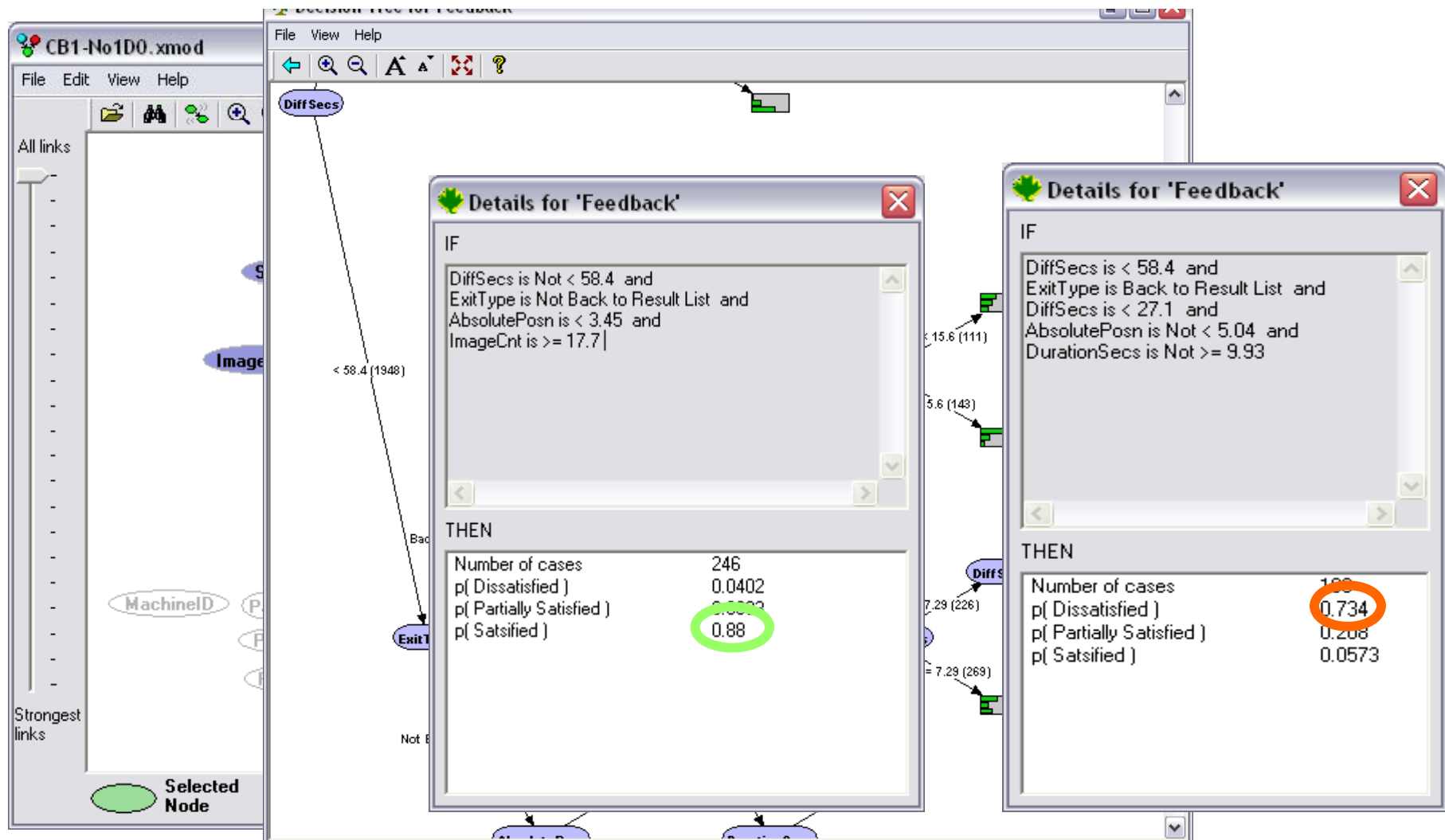
Result-Level	Session-Level
Diff Secs, Duration Secs	Averages of result-level measures (Dwell Time and Position)
Scrolled, ScrollCnt, AvgSecsBetweenScroll, TotalScrollTime, MaxScroll	Query count
TimeToFirstClick, TimeToFirstScroll	Results set count
Page, Page Position, Absolute Position	Results visited
Visits	End action
Exit Type	
ImageCnt, PageSize, ScriptCnt	
Added to Favorites, Printed	

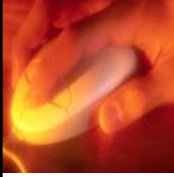




Data Analysis, cont'd

[Fox et al., 2003]





Result-Level Findings

[Fox et al., 2003]

1. Dwell time, clickthrough and exit type strongest predictors of SAT
2. Printing and Adding to Favorites highly predictive of SAT when present
3. Combined measures predict SAT better than clickthrough





Result Level Findings, cont'd

[Fox et al., 2003]

Feedback	Num	Percent
Satisfied	1481	0.38
Partially Satisfied	1147	0.30
Dissatisfied	1055	0.27
Could not evaluate	172	0.04
Grand Total	3855	

Only clickthrough

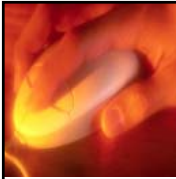
PageResultFB TestData (g/t)				
	Sat	PSat	DSat	
Sat	172	53	0	0.7
PSat	67	91	36	0.47
DSat	39	86	134	0.52
				0.57 correct
				0.92 one-off

Combined measures

PageResultFB TestData (t/g) - Conf > 0.5				
	Sat	PSat	DSat	
Sat	152	33	2	0.772
PSat	25	26	5	0.464
DSat	9	56	89	0.578
				0.656 correct
				0.948 one-off

Combined measures with confidence of > 0.5 (80-20 train/test split)





Learning Result Preferences in Rich User Interaction Space

[Agichtein et al., 2006]

- Observed and Distributional features
 - Observed features: aggregated values over all user interactions for each query and result pair
 - Distributional features: deviations from the “expected” behavior for the query
- Represent user interactions as vectors in “Behavior Space”
 - **Presentation**: what a user sees *before* click
 - **Clickthrough**: frequency and timing of clicks
 - **Browsing**: what users do *after* the click





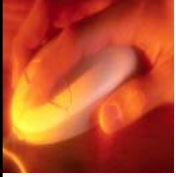
Features for Behavior Representation

[Agichtein et al., SIGIR2006]

<i>Presentation</i>	
ResultPosition	Position of the URL in Current ranking
QueryTitleOverlap	Fraction of query terms in result Title
<i>Clickthrough</i>	
DeliberationTime	Seconds between query and first click
ClickFrequency	Fraction of all clicks landing on page
ClickDeviation	Deviation from expected click frequency
<i>Browsing</i>	
DwellTime	Result page dwell time
DwellTimeDeviation	Deviation from expected dwell time for query

Sample Behavior Features



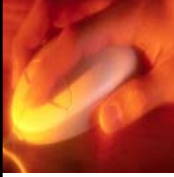


Predicting Result Preferences

[Agichtein et al., SIGIR2006]

- Task: predict pairwise preferences
 - A **judge** will prefer Result A > Result B
- Models for preference prediction
 - Current search engine ranking
 - Clickthrough
 - Full user behavior model





User Behavior Model

[Agichtein et al., SIGIR2006]

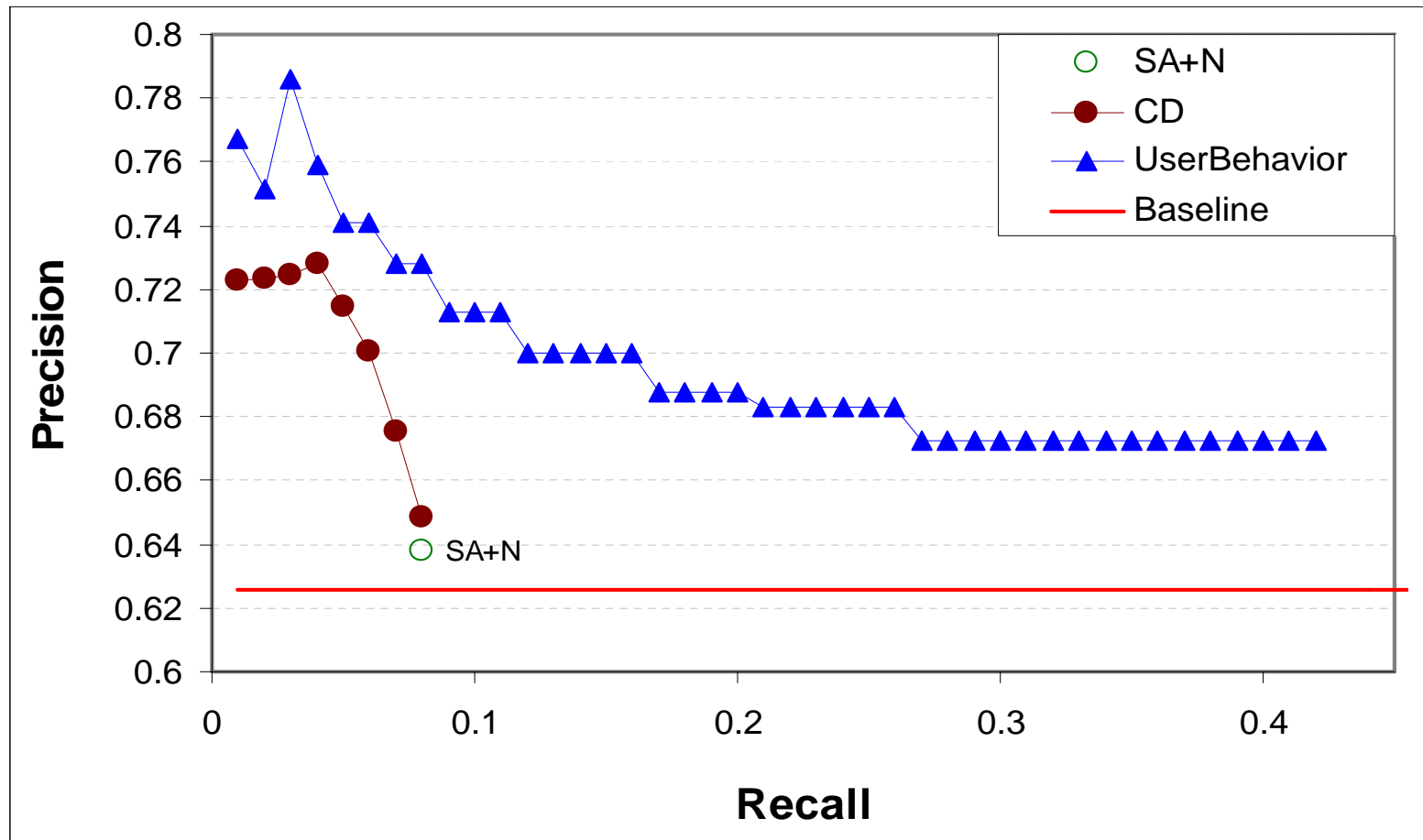
- Full set of interaction features
 - Presentation, clickthrough, browsing
- **Train** the model with explicit judgments
 - Input: behavior feature vectors for each query-page pair in rated results
 - Use **RankNet** (Burges et al., [ICML 2005]) to discover model weights
 - Output: a neural net that can assign a “relevance” score to a behavior feature vector





Results: Predicting User Preferences

[Agichtein et al., SIGIR2006]



- Baseline < SA+N < CD << UserBehavior
- Rich user behavior features result in dramatic improvement



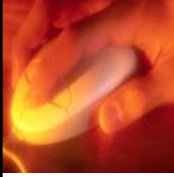


Observable Behavior

Minimum Scope

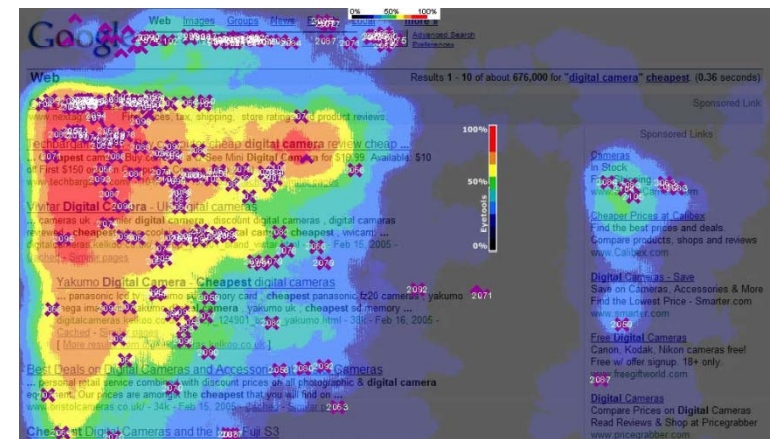
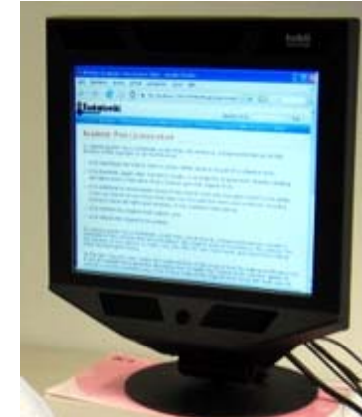
Behavior Category	Segment	Object	Class
	Examine		
	Retain	Print	Bookmark Save Purchase Delete
	Reference	Copy / paste Quote	Forward Reply Link Cite
	Annotate	Mark up	Rate Publish Organize

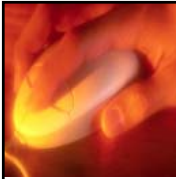




Eye Tracking

- Unobtrusive
- Relatively precise (accuracy: 1° of visual angle)
- Expensive
- Mostly used as „passive“ tool for behavior analysis, e.g. visualized by heatmaps:
- We use eye tracking for immediate implicit feedback taking into account temporal fixation patterns

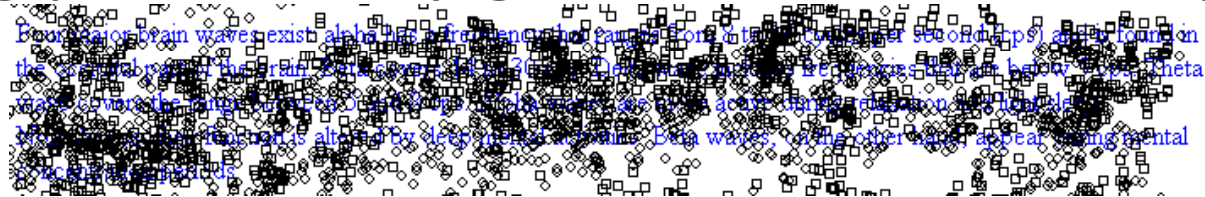




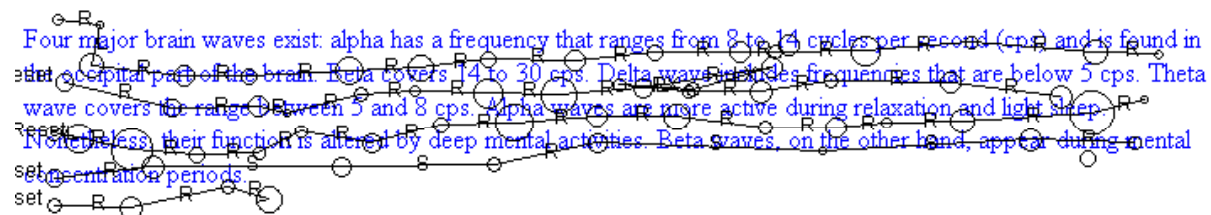
Using Eye Tracking for Relevance Feedback

[Buscher et al., 2008]

- Starting point: Noisy gaze data from the eye tracker.



2. Fixation detection and saccade classification

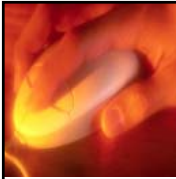


3. Reading (red) and skimming (yellow) detection line by line

Four major brain waves exist: alpha has a frequency that ranges from 8 to 14 cycles per second (cps) and is found in the occipital part of the brain. Beta covers 14 to 30 cps. Delta wave includes frequencies that are below 5 cps. Theta wave covers the range between 5 and 8 cps. Alpha waves are more active during relaxation and light sleep. Nonetheless, their function is altered by deep mental activities. Beta waves, on the other hand, appear during mental concentration periods.

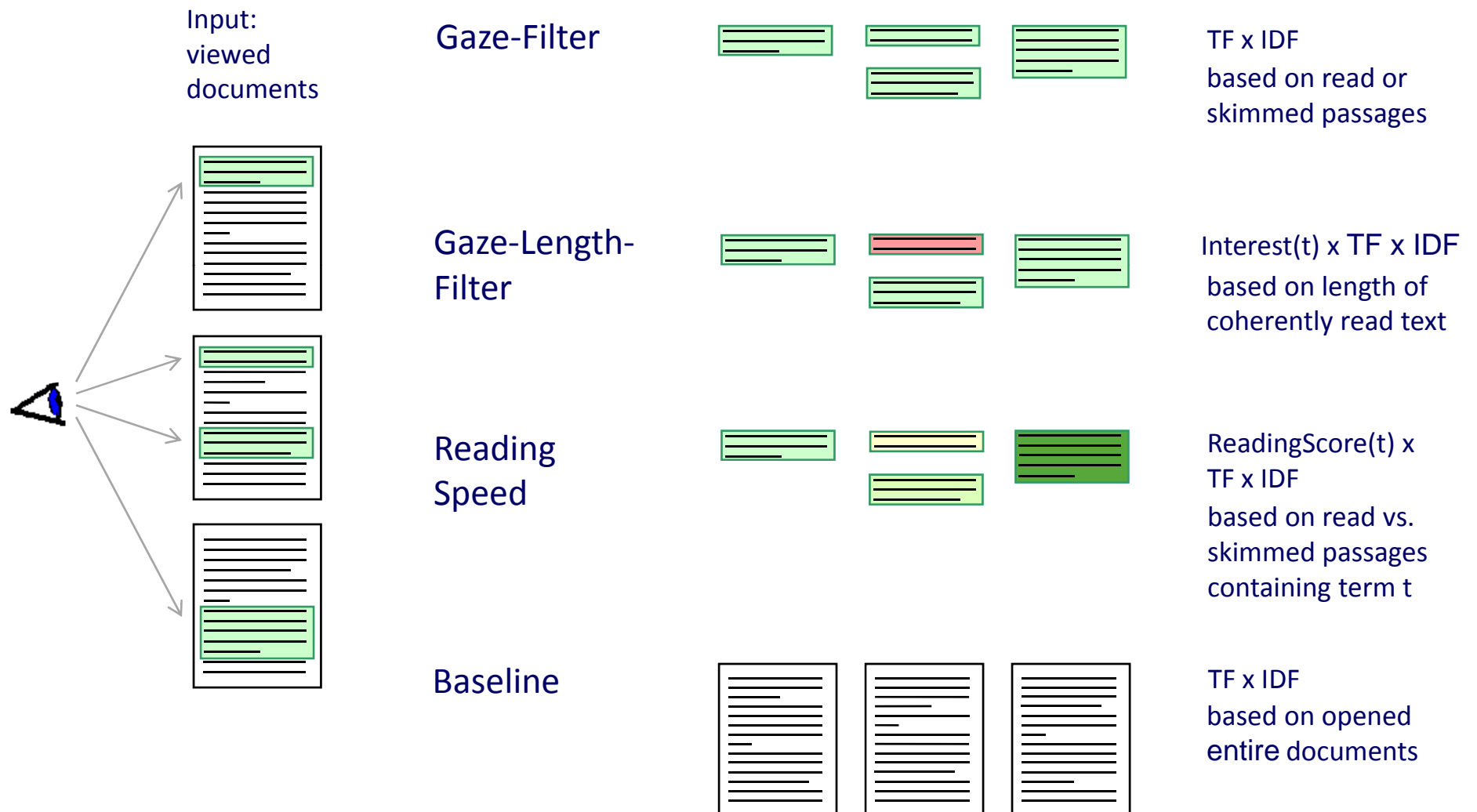
See G. Buscher, A. Dengel, L. van Elst: "Eye Movements as Implicit Relevance Feedback", in CHI '08

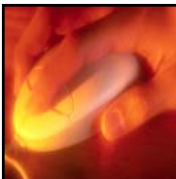




Three Feedback Methods Compared

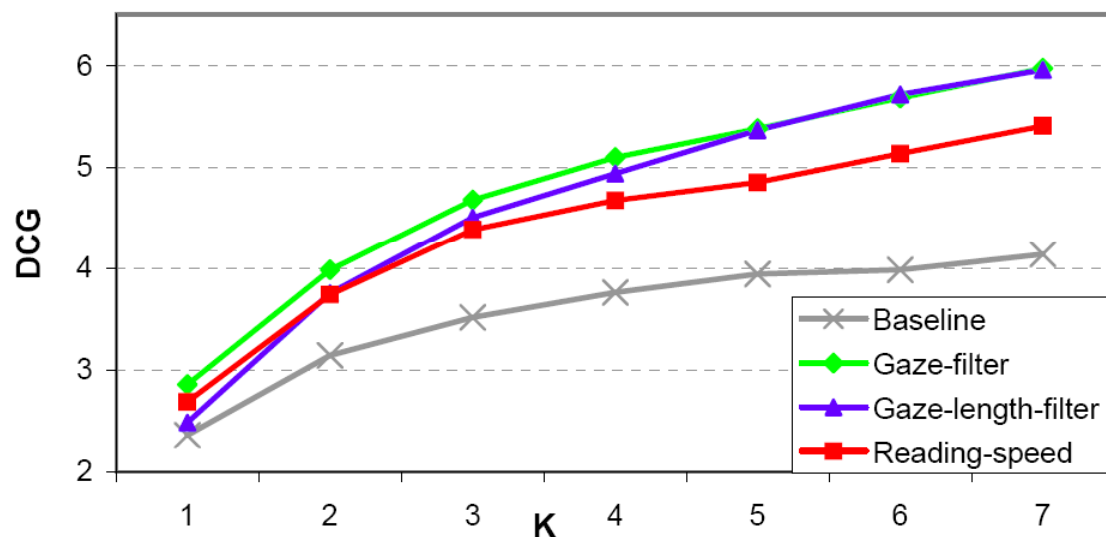
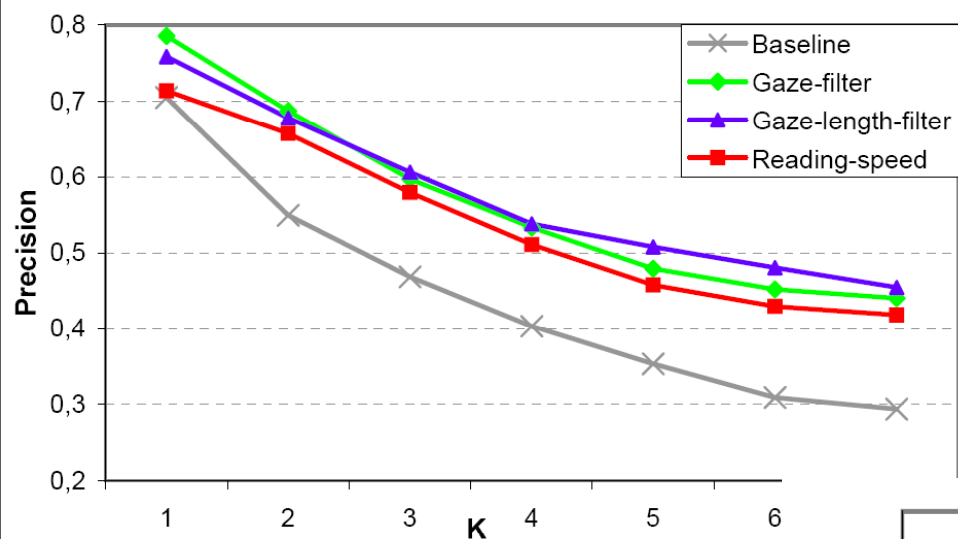
[Buscher et al., 2008]





Eye-based RF Results

[Buscher et al., 2008]



You can try this too...

- Competition: “Inferring relevance from eye movements”
 - Predict relevance of titles, given the eye movements.
 - 11 participants, best accuracy 72.3% (TU Graz)
- **Data available at:**
<http://www.cis.hut.fi/eyechallenge2005/>
- **Workshop** on held Machine Learning for Implicit Feedback and User Modeling at NIPS'05



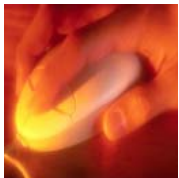
Lecture 2 Summary



- Explicit Feedback in IR
 - Query expansion
 - User control



- From Clicks to Relevance



- 3. Rich Behavior Models
 - + Browsing
 - + Session/Context information
 - + Eye tracking



Key References and Further Reading

- Marti **Hearst**, Search User Interfaces, 2009, Chapter 6 “Query Reformulation”: <http://searchuserinterfaces.com/>

Kelly, D. and Teevan, J. *Implicit feedback for inferring user preference: a bibliography*. SIGIR Forum 37, 2 (Sep. 2003)

Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. Accurately interpreting clickthrough data as implicit feedback., SIGIR 2005

Agichtein, E., Brill, E., Dumais, S., and Ragno, R. *Learning user interaction models for predicting web search result preferences*, SIGIR 2006

Buscher, G., Dengel, A., and van Elst, L. *Query expansion using gaze-based feedback on the subdocument level.*, SIGIR 2008

Chapelle, O, and Y. Zhang, A Dynamic Bayesian Network Click Model for Web Search Ranking, WWW 2009

Piwowarski, B, Dupret, G, Jones, R: *Mining user web search activity with layered bayesian networks or how to capture a click in its context*, WSDM 2009

