The Role of Information Scent in On-line Browsing:

Extensions of the ACT-R Utility and Concept Formation Mechanisms

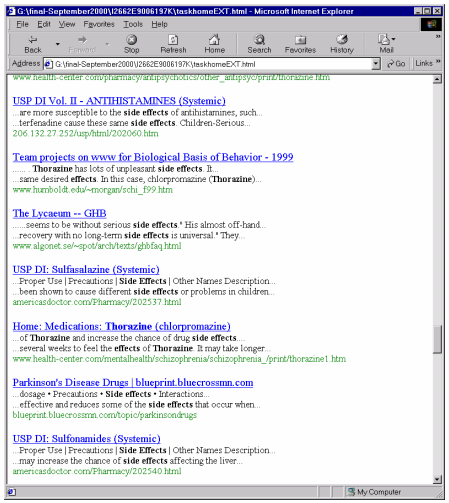
Peter Pirolli

User Interface Research Area

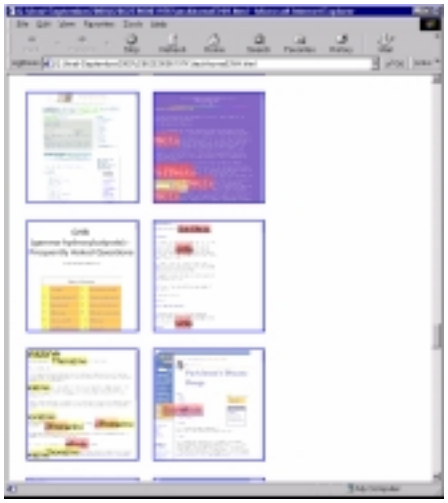
Supported in part by
Office of Naval Research (ONR)
Advanced Research and Development Agency (ARDA)

parc
Palo Alto Research Center

Information Scent

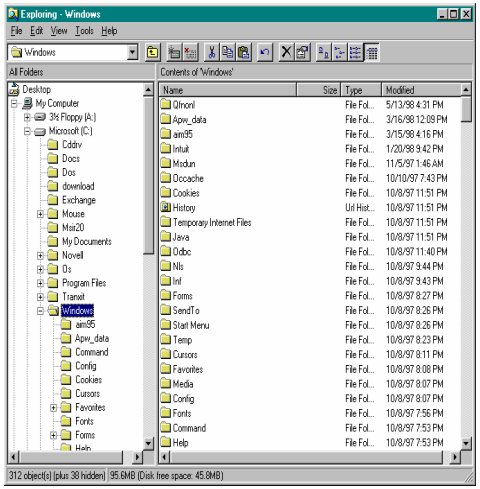


(a)



(b)

(c)



(d)

Browsing often requires that we use visual or textual cues to direct our actions.

These local (proximal) cues are called *information scent*.

The *theory of information scent* is a psychological theory

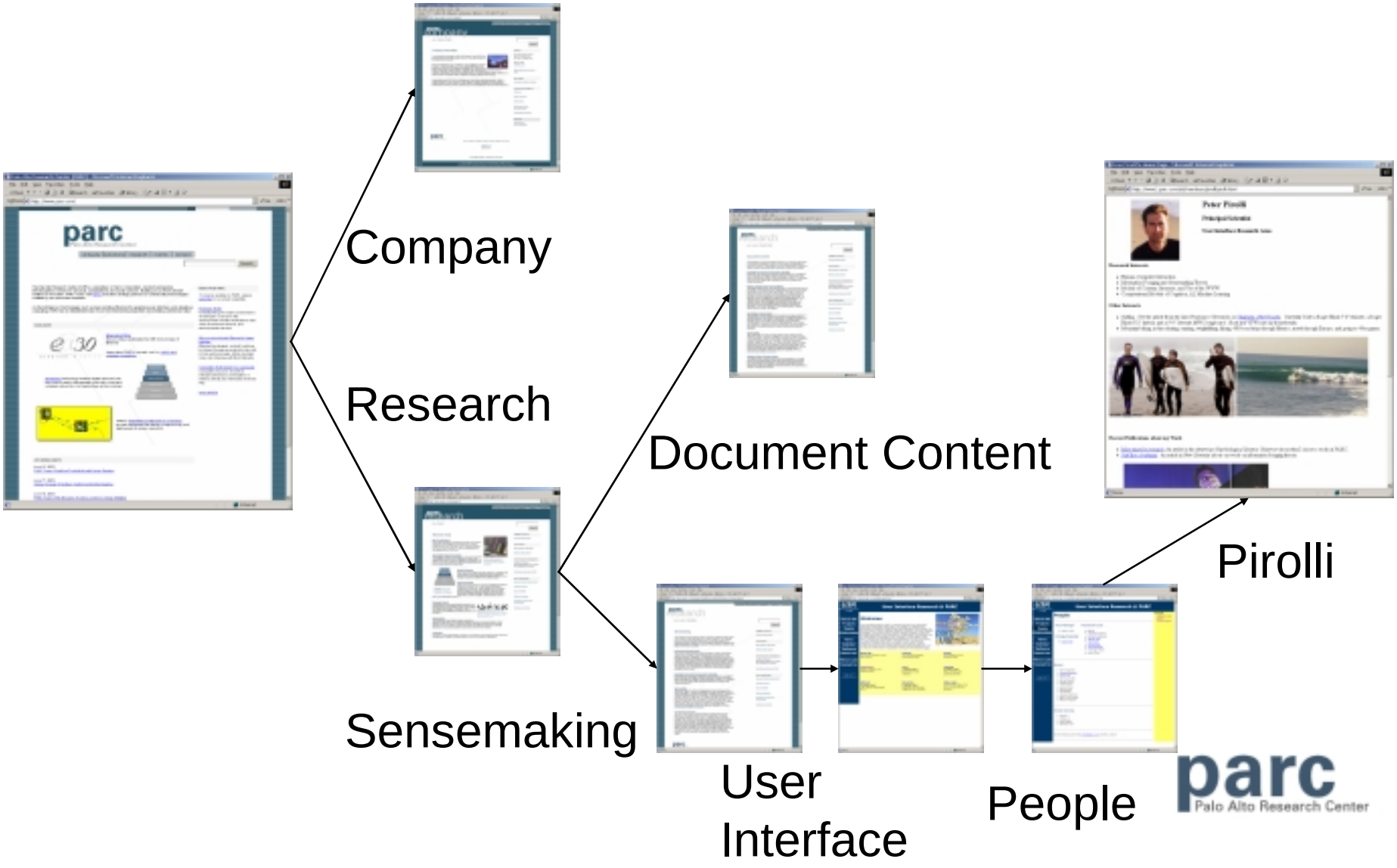
Used to develop new user interface designs and new Web site evaluation tools

parc
Palo Alto Research Center

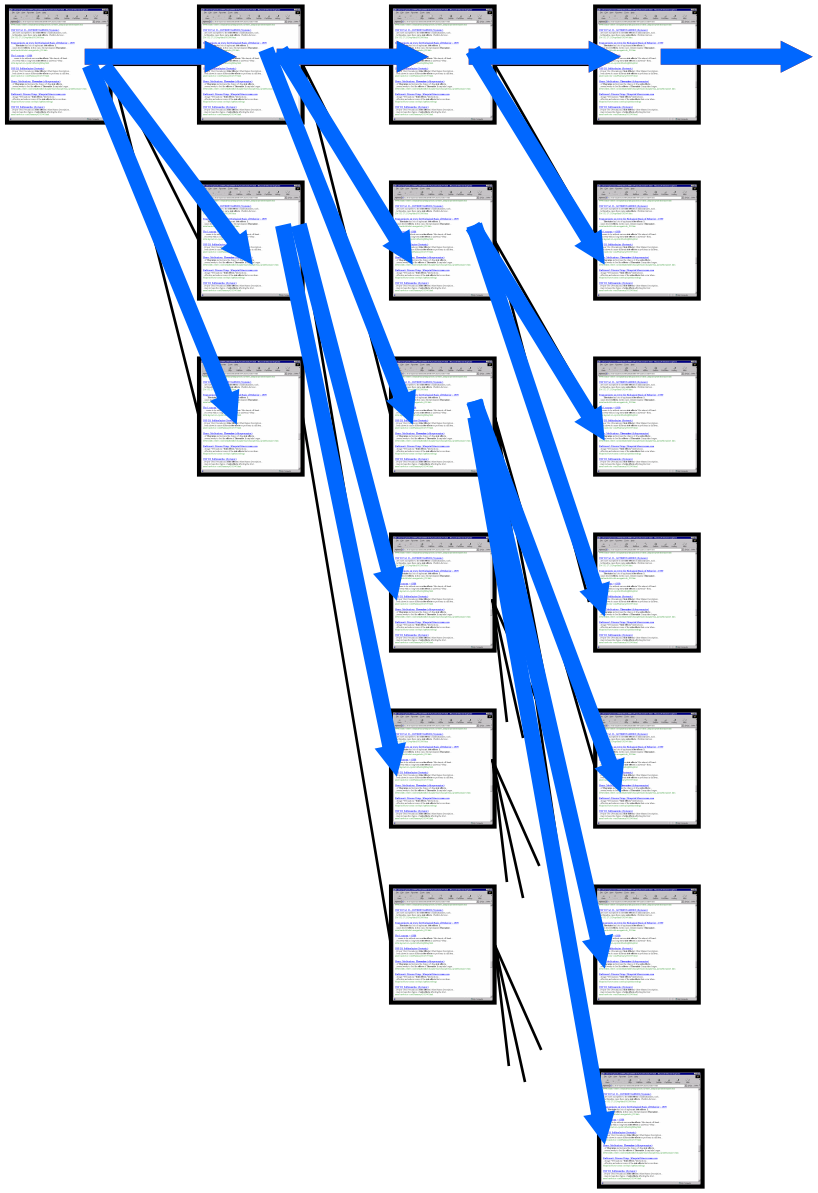
Overview

- Why information scent is important
- SNIF-ACT Model of Navigation
  - Where to go next (navigational choices)?
  - When to stop (leaving a Web site)?
- InfoCLASS Model of topic category formation
  - What categories of topics are where (learning about the information environment)

Example: Looking for Pirolli's Personal Page on the PARC Site



Hierarchical Navigation Spaces

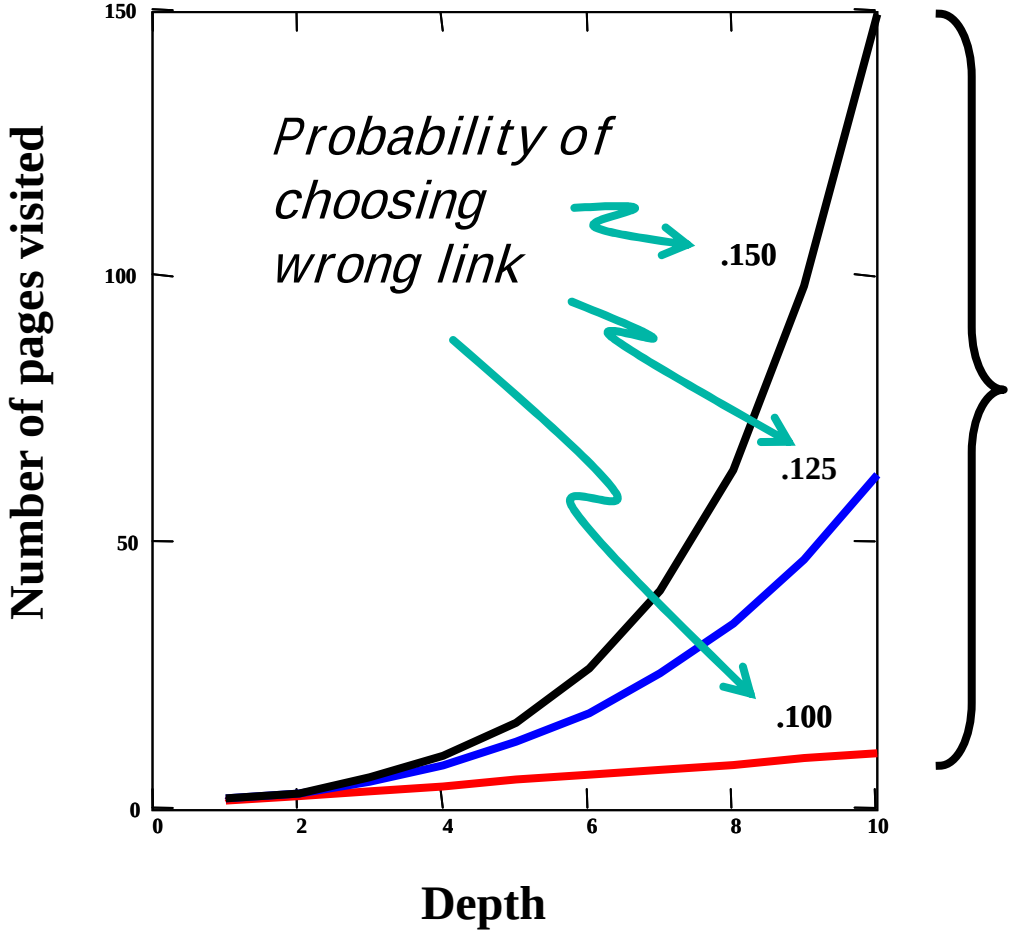


Strong Information Scent
= Navigation Cost
Linear in Depth

Weak Information Scent
= Navigation Cost
Exponential in Depth

parc
Palo Alto Research Center

Navigation Costs as Function of Information Scent

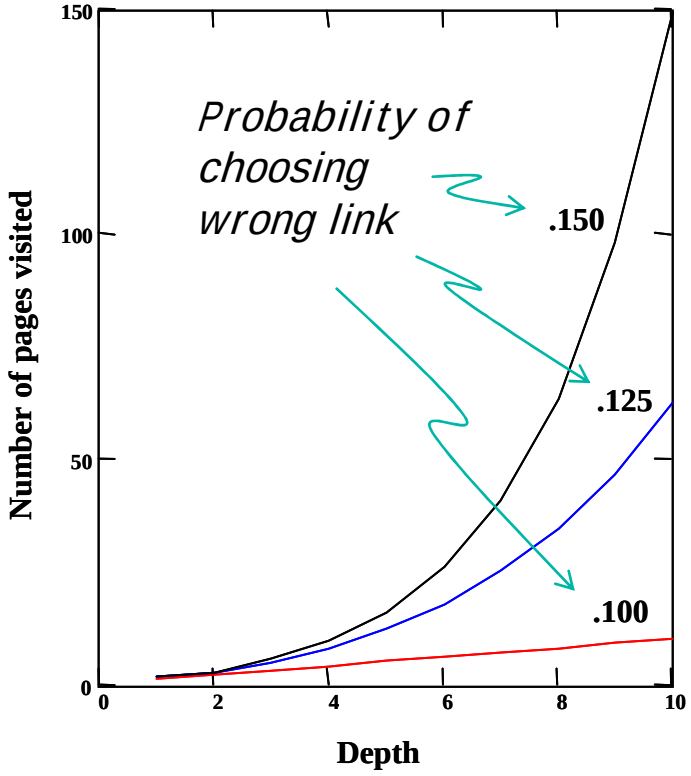


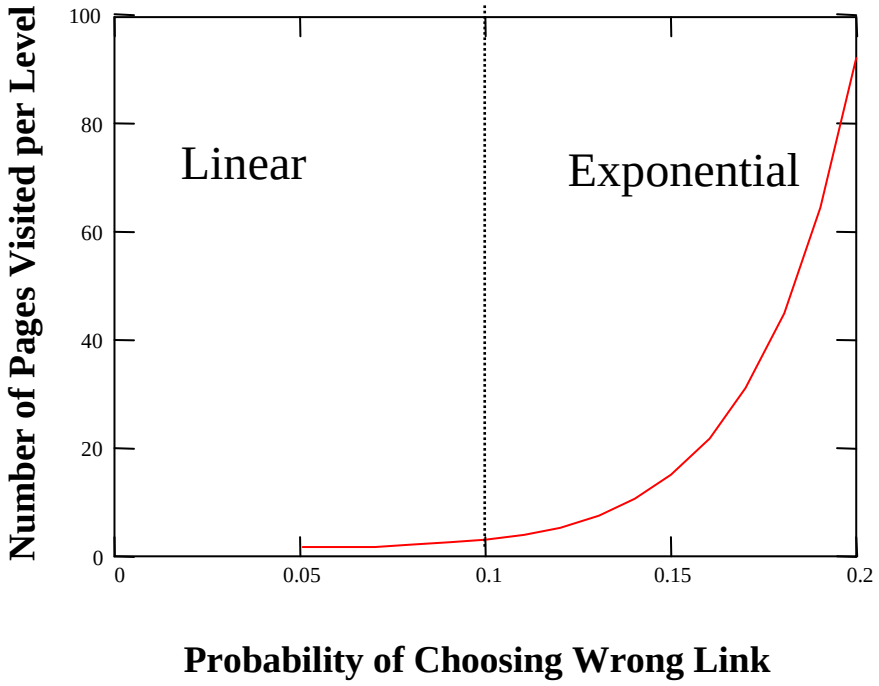
Probabilities within range of empirical values in Woodruff et al. (2002)

Notes: Average branching factor = 10
Depth = 10

parc
Palo Alto Research Center

Phase Transition in Navigation Costs as Function of Information Scent

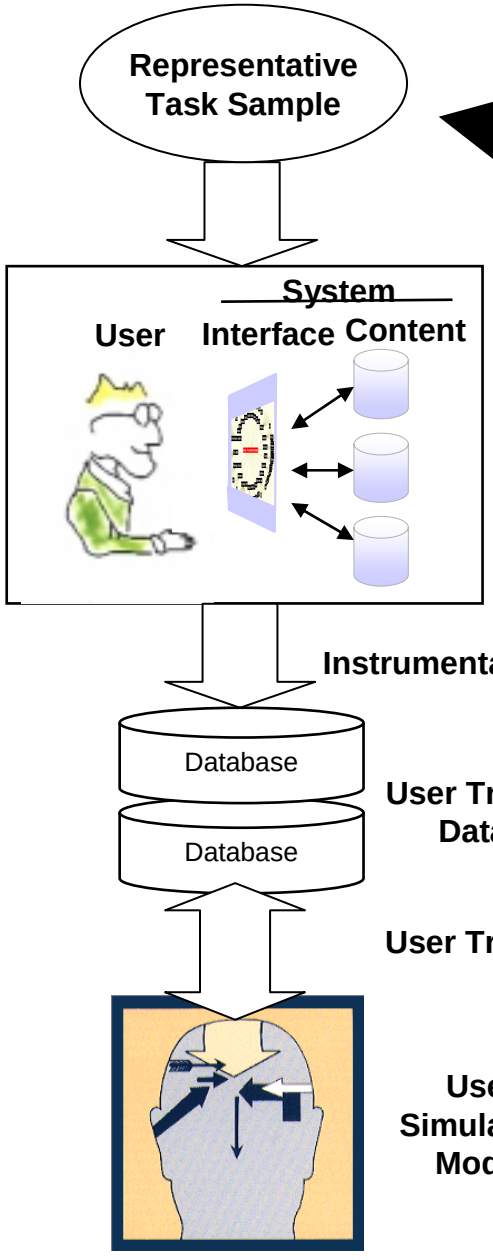




Notes: Average branching factor = 10
Depth = 10

parc
Palo Alto Research Center

WWW Studies



Survey of user tasks (N = 2188). Develop *ecologically valid* tasks for study in experiments.

Experiments

Develop instrumentation to record user trace data.

Analyze data. Develop cognitive model. Validate against user trace data.

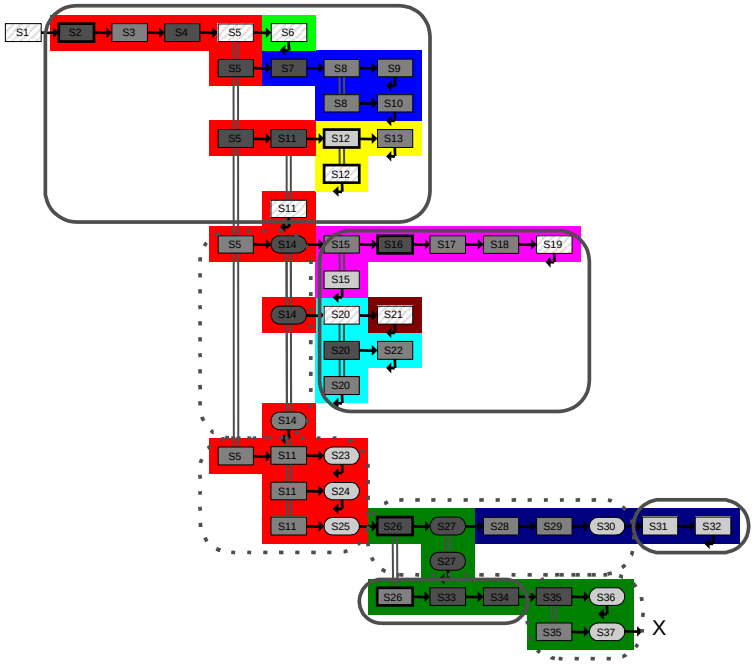
WWW Experiment & Model

- 12 Stanford University students
- 6 tasks
- 2 tasks analyzed and modeled for 4 participants
- Example tasks
  - Find specific movie posters for your new living room
  - Find dates for a performance by a comedy troupe

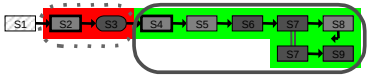
Pirolli, P. & Fu, W. (2003). SNIF-ACT: A model of information foraging on the World Wide Web. Proceedings of the Conference on User Modeling.

Web Behavior Graphs (WBGs)

Find a poster for the movie "Antz"



Find the tour schedule of the "Second City Comedy Troupe"



- Links providing better information scent yield more direct navigation
- People abandon Web sites when information scent diminishes

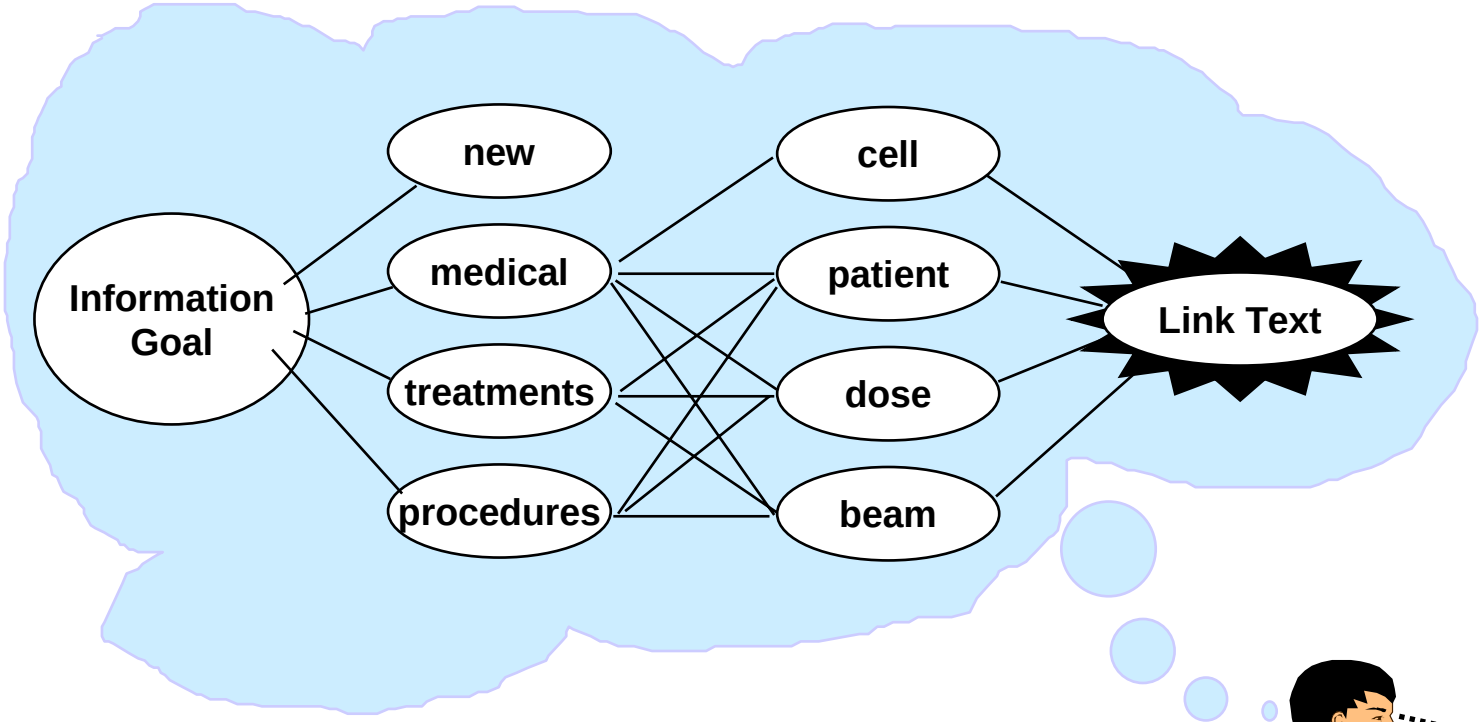
Card, S., Pirolli, P., Van Der Wege, M., Morrison, J., Reeder, R., Schraedley, P., & Boshart, J. (2001). ***CHI Proceedings***

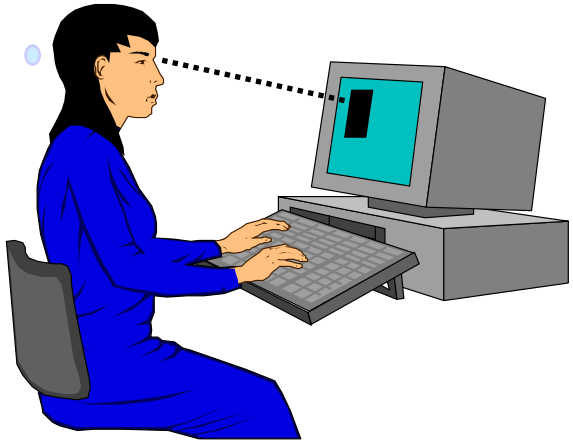
SNIF-ACT

- Scent-based Navigation and Information Foraging in the ACT theory
- Declarative knowledge
  - User goal (e.g. Find the poster for “Antz”)
  - Perceived aspects of Web page & browser
  - Large spreading activation network representing word associations
- Procedural knowledge
  - Productions representing basic Web browsing actions
- Utility: Information Scent
  - Mutual relevance between link text and user goal

parc
Palo Alto Research Center

Cognitive Model of Information Scent





- *Spreading activation*
  - Activation reflects likelihood of relevance given past history and current context

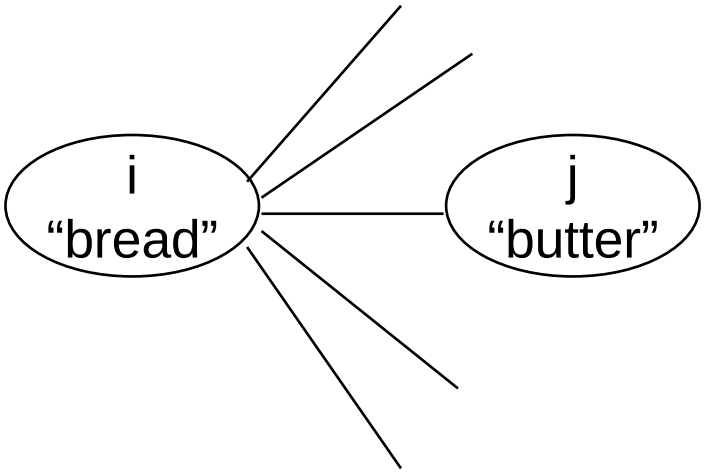
Spreading activation

Activation of node i

$$A_i = B_i + \Sigma W_j S_{ji}$$

Base-level activation

Activation spread from linked nodes j



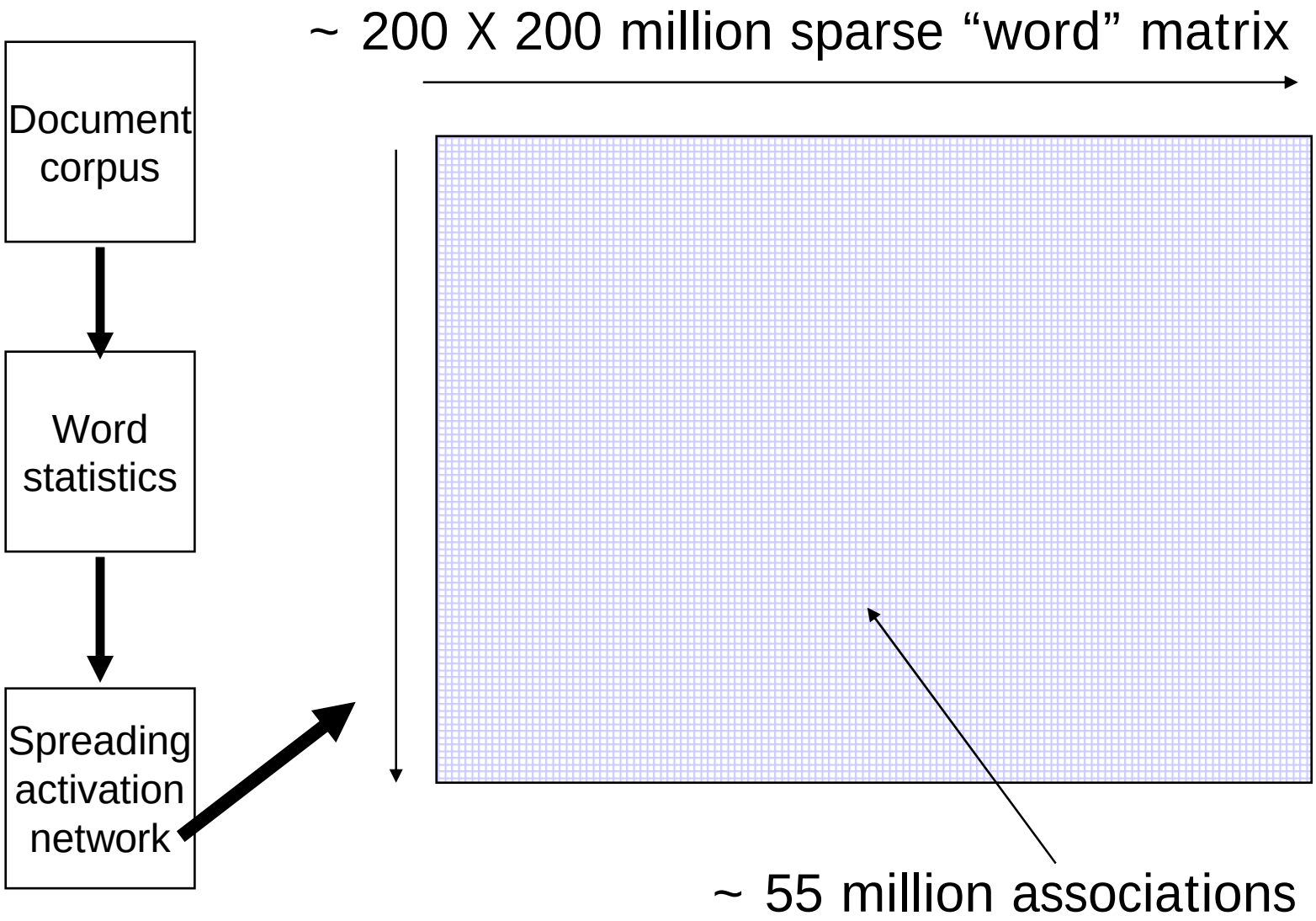
Base-level reflects log odds of occurrence

$$B_i = \ln\left(\frac{\Pr(i)}{\Pr(\text{not } i)}\ \psi\right)$$

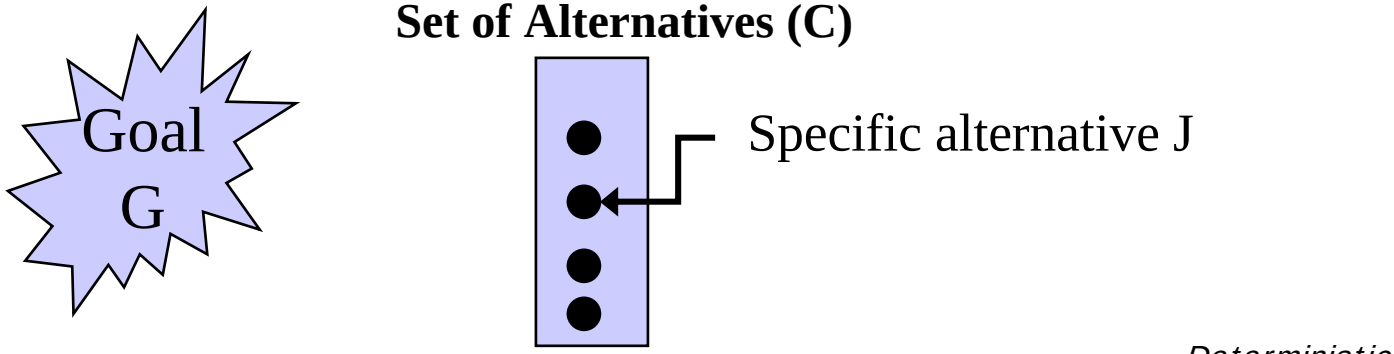
Strength of link spread reflects log likelihood odds of cooccurrance

$$S_{ji} = \ln\left(\frac{\Pr(j|i)}{\Pr(j|\text{not } i)}\ \psi\right)$$

spreading activation networks



Information Scent:
Random Utility Model based on Activation



Utility of alternative J, out of set C, in context of goal G:

$$U_{J|G} = V_{J|G} + \varepsilon_{J|G}$$

Deterministic

Stochastic

Deterministic utility is based on activation:

$$V_{J|G} = \sum_{i \in G} A_i$$

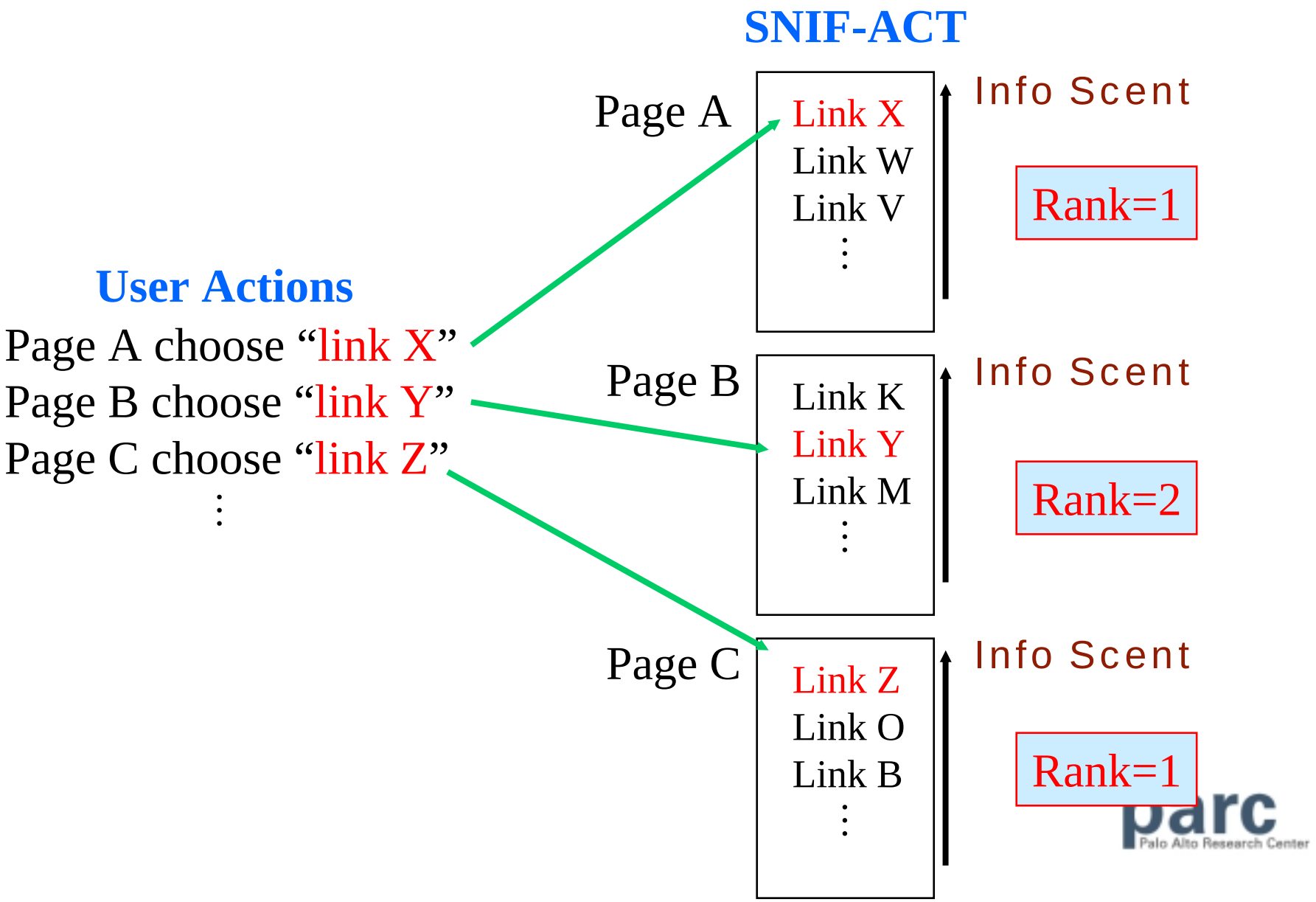
Summed activation from scent cues to goal

Probability of choosing alternative J:

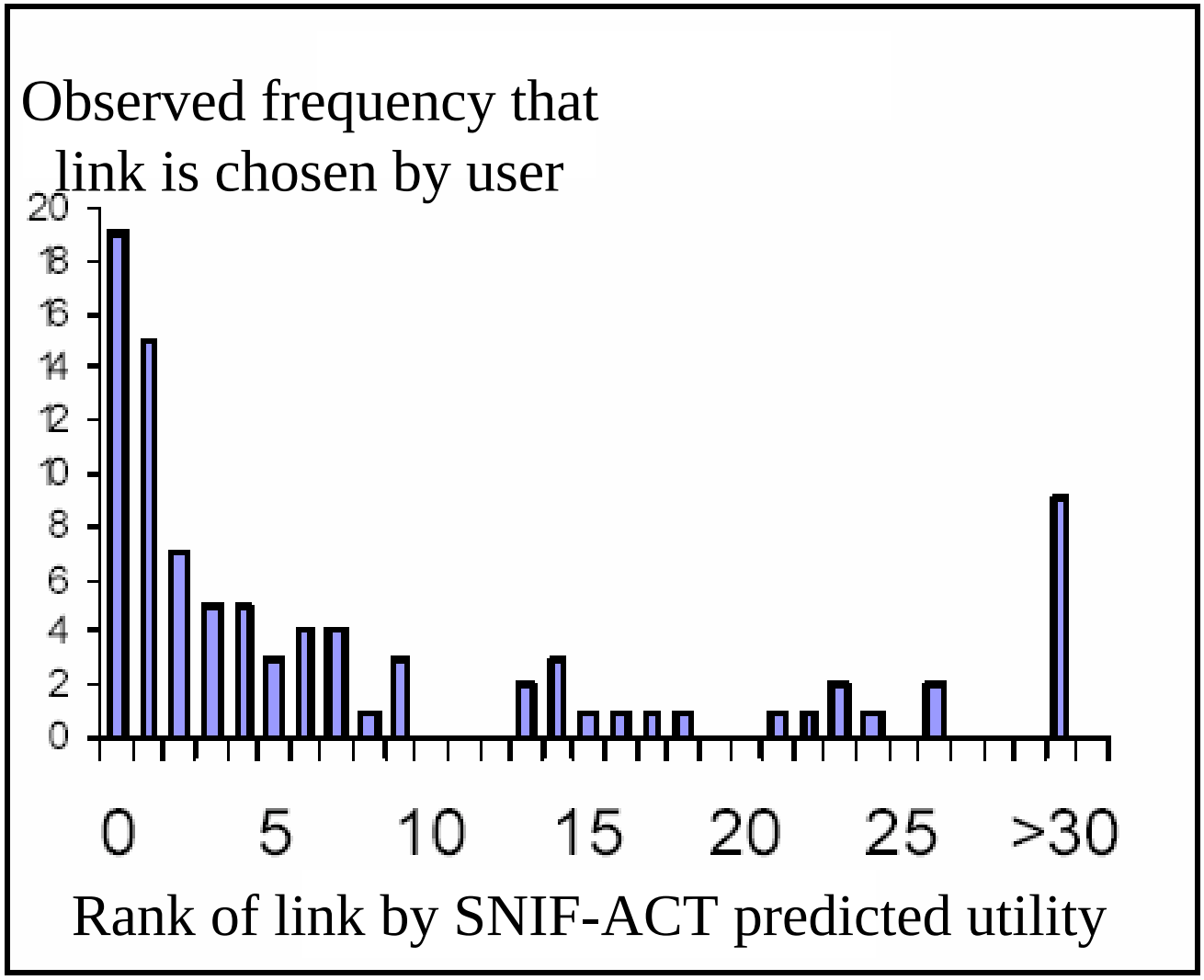
$$\Pr(J \mid G, C) = \frac{e^{\mu V_{J|Q}}}{\sum_{K \in C} e^{\mu V_{K|Q}}}$$

parc
Palo Alto Research Center

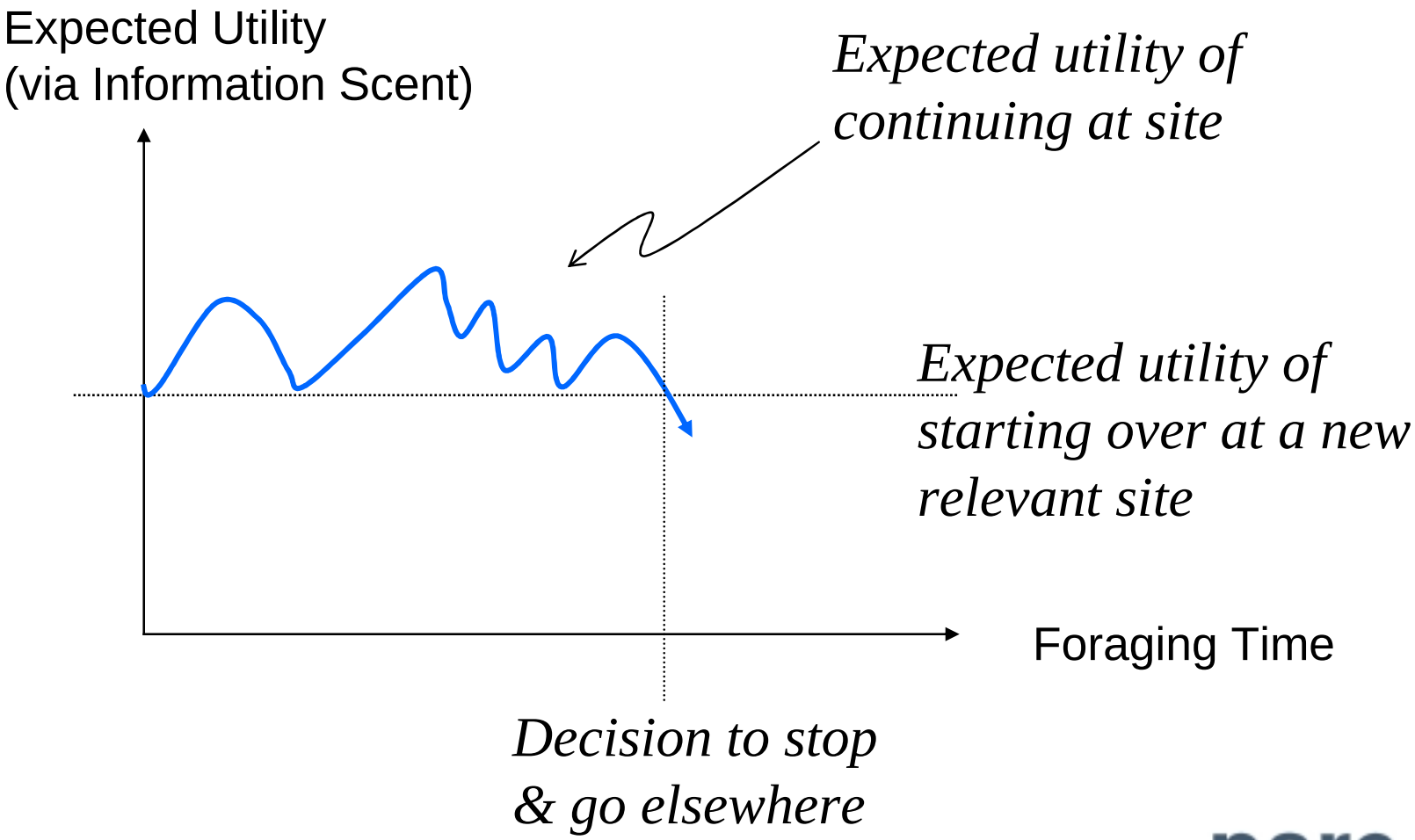
Evaluation:Link-following actions



Observed vs Predicted Link Choice

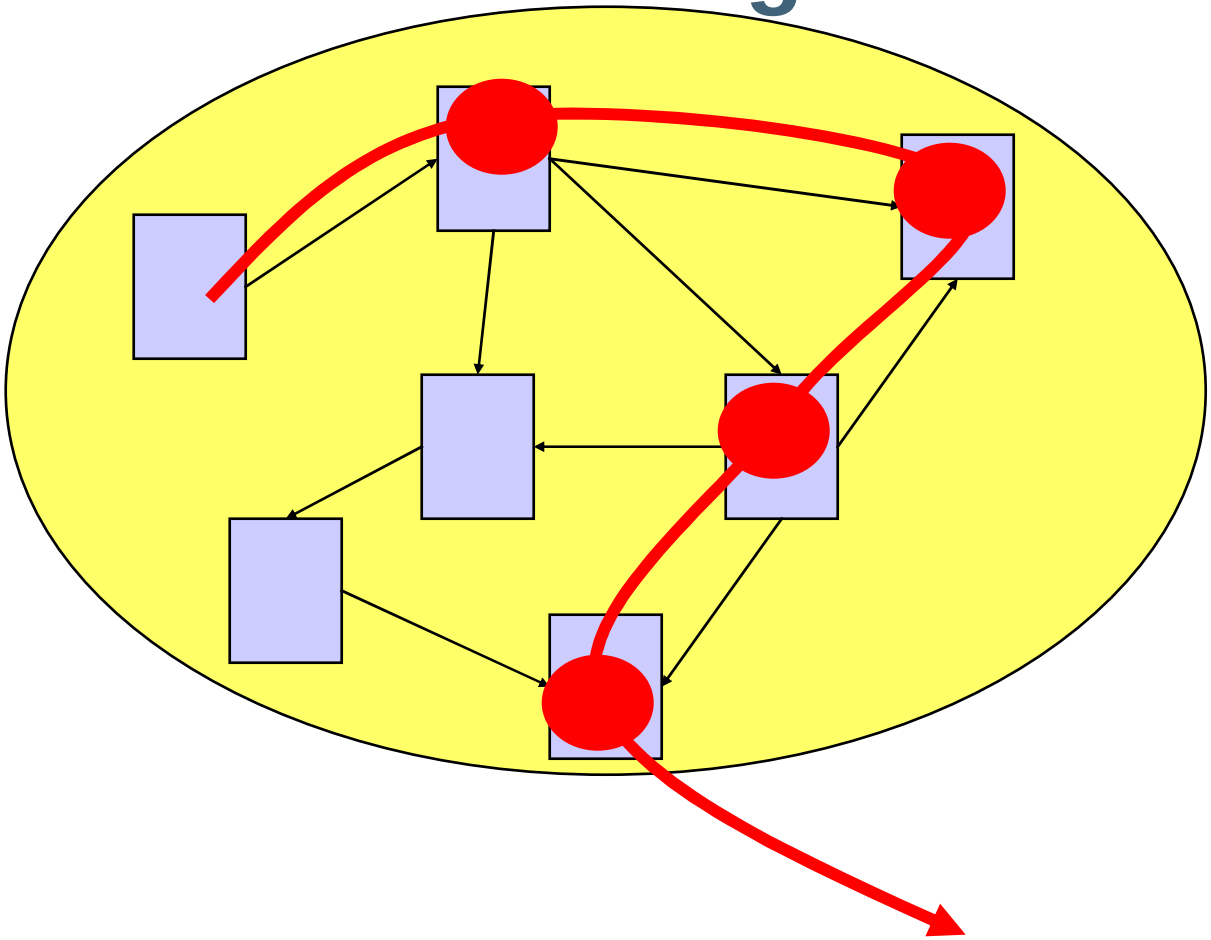


Theory: Decision to Leave Web Site



parc
Palo Alto Research Center

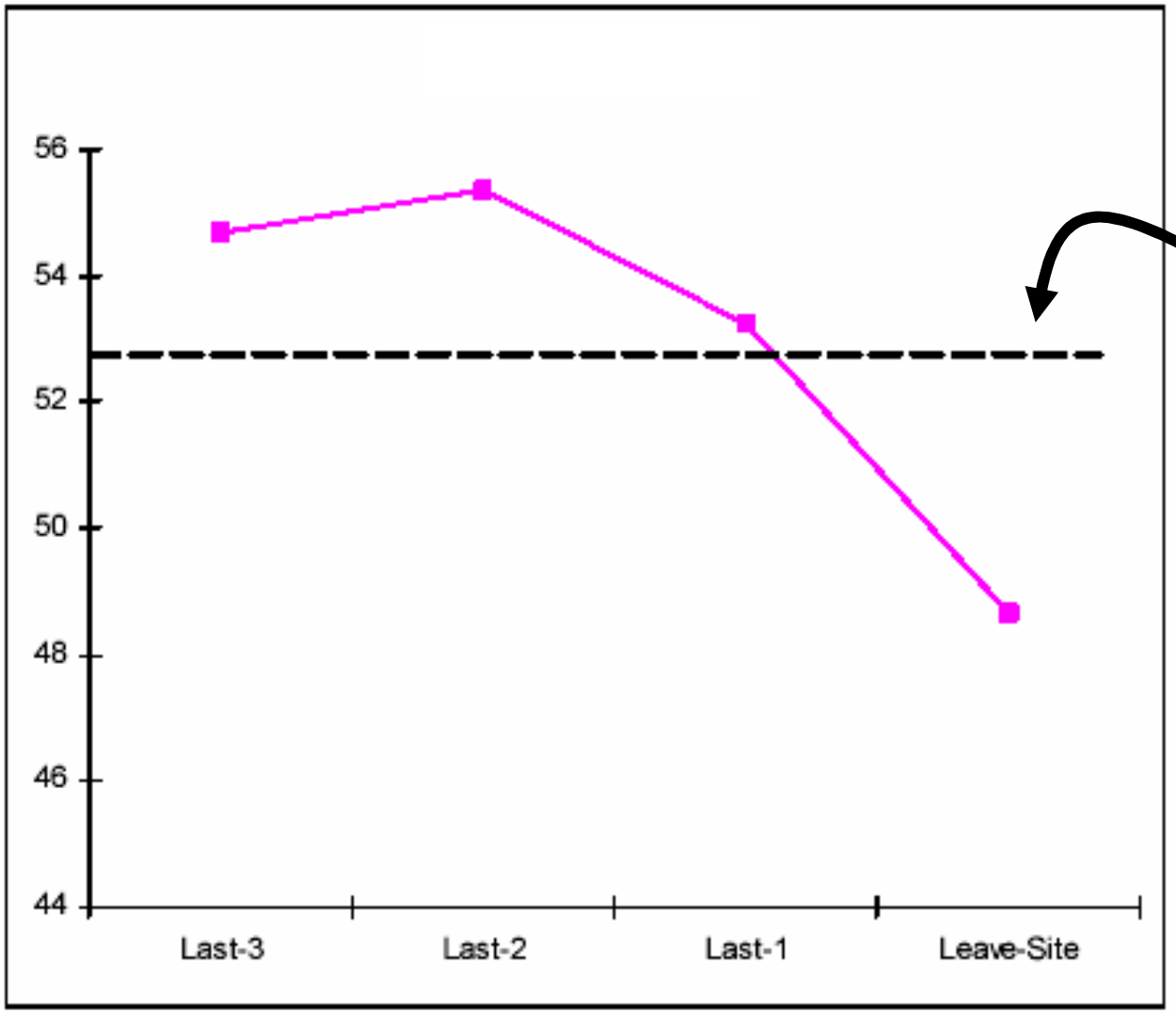
Data: Sequences of Moves Just Prior to Leaving a Site



parc
Palo Alto Research Center

Match of SNIF-ACT to User Data

SNIF-ACT Predicted utility of next link



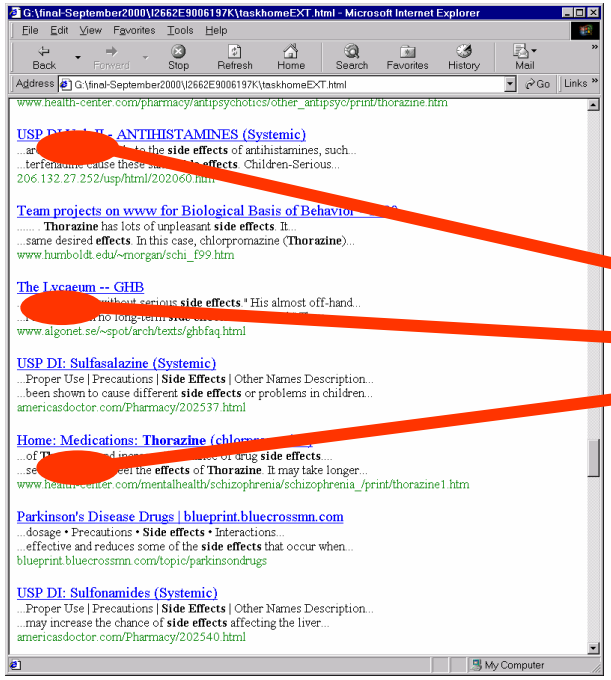
SNIF-ACT Average utility of new site

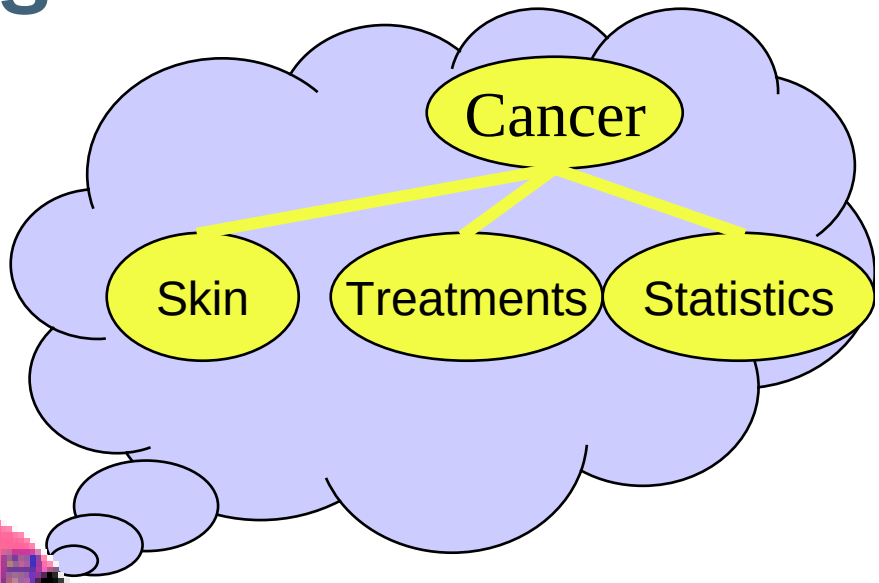
User Move Before Leaving Site

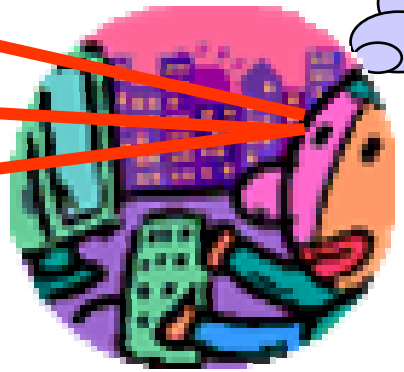
parc
Palo Alto Research Center

Forming Concepts about Information Sources

Standard Web Site Search Engine



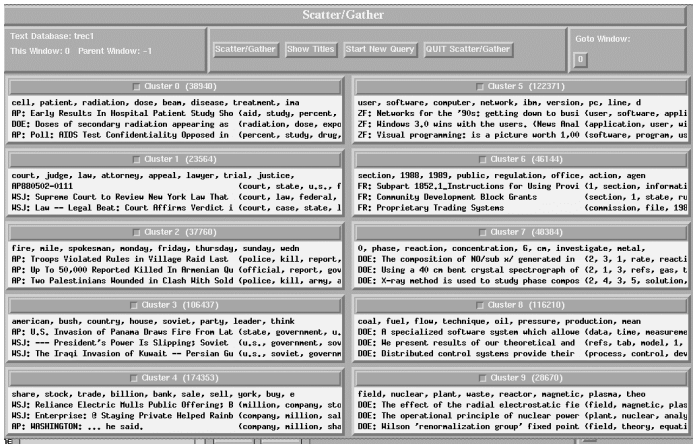


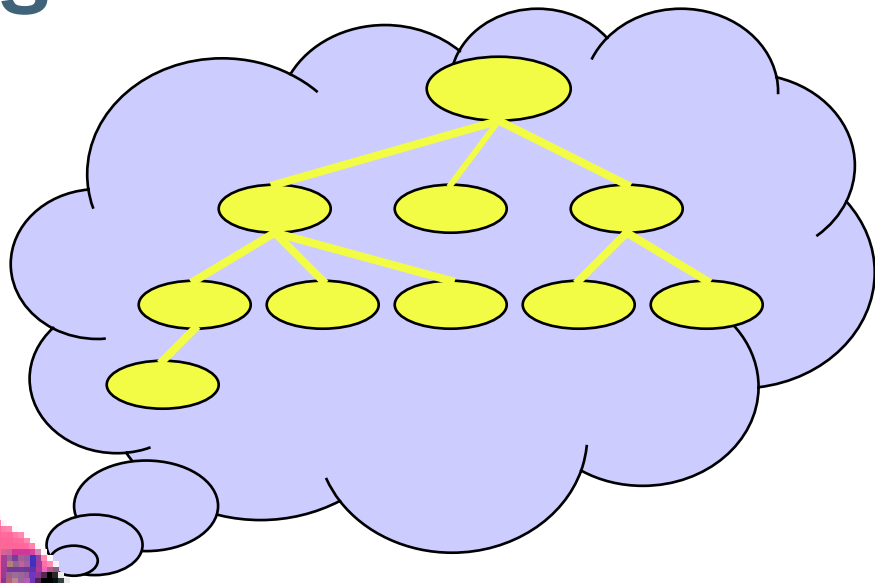


parc
Palo Alto Research Center

Forming Concepts about Information Sources

Scatter/Gather
Document Clustering Browser





parc
Palo Alto Research Center

Browser Study

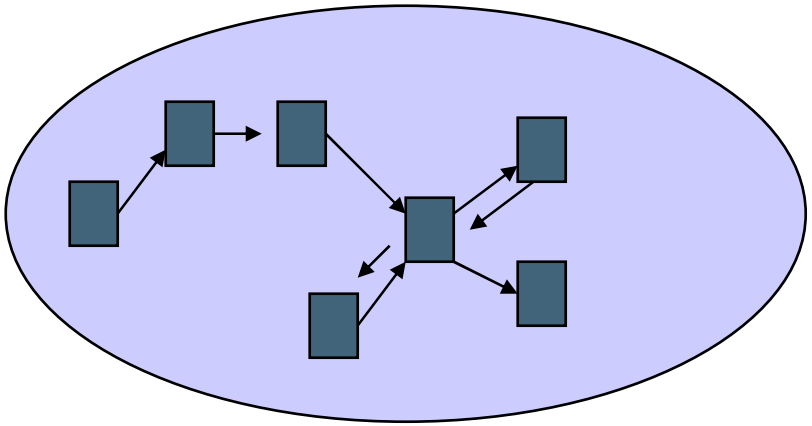
- Browsers
  - Scatter-Gather (*N* = 8)
  - Standard Similarity Search Engine (*N* = 8)
- 12 Tasks (TREC)
  - E.g., Find documents on new medical procedures for cancer

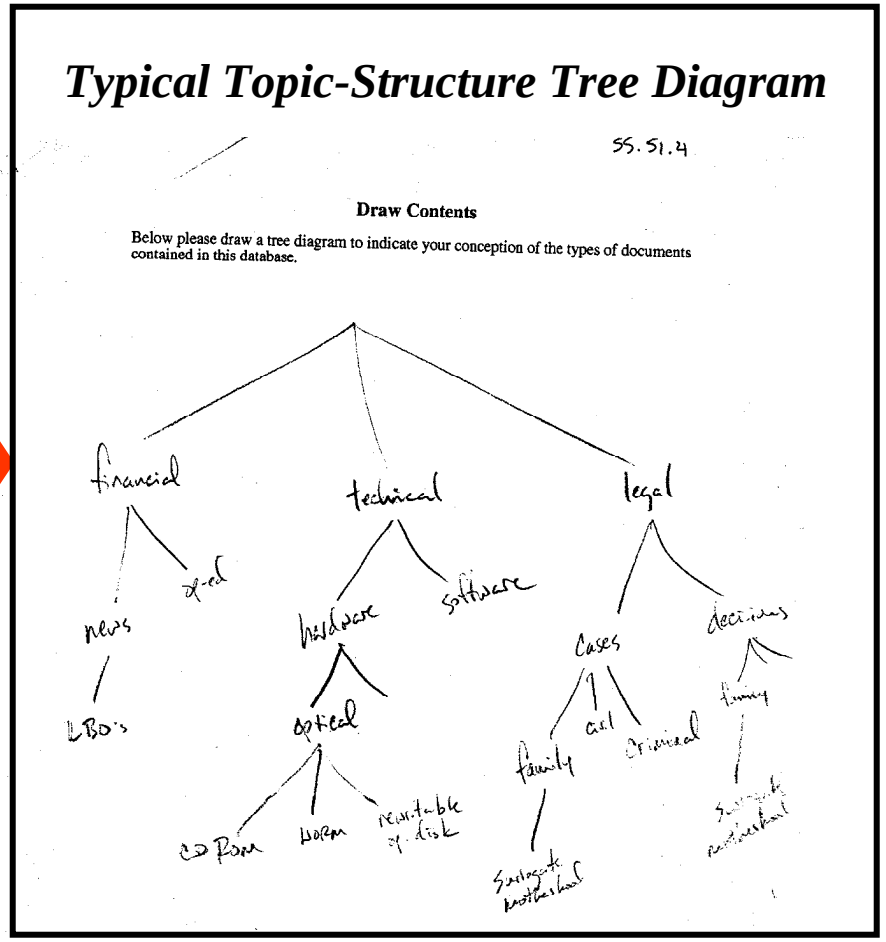
Pirolli, P., Hearst, M., Schank, P., & Diehl, C. (1996). Scatter-Gather browsing communicates the topic structure of a very large text collection. *Proceedings of the 'CHI '96 Conference*.

parc
Palo Alto Research Center

Inferred Topic Structure

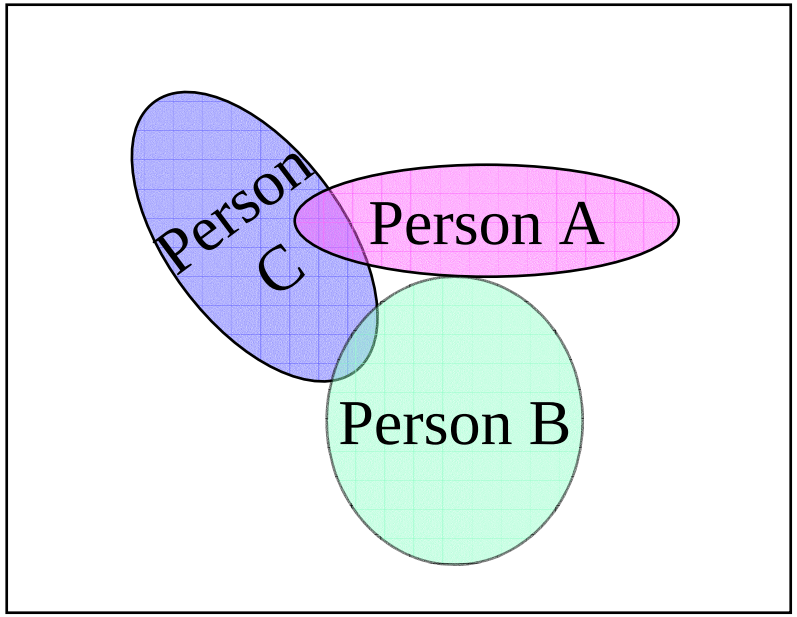
Browse documents



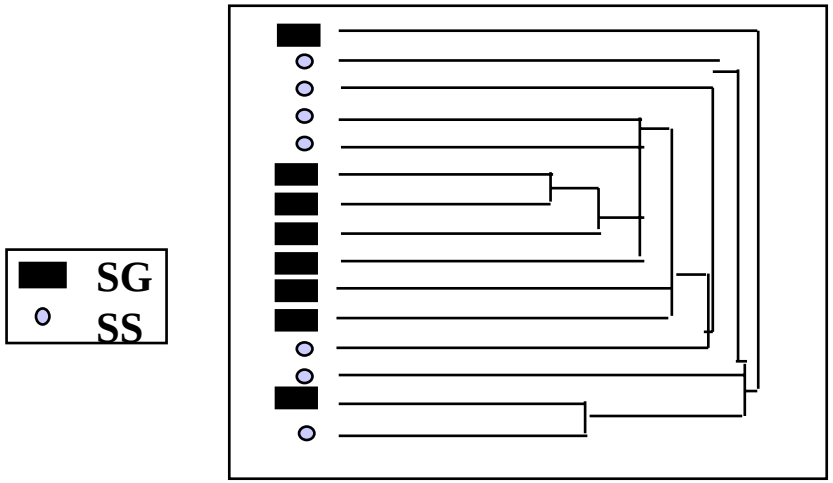


parc
Palo Alto Research Center

Coherence of Mental Topics Among Users



Hierarchical clustering of users based on similarity of topic trees



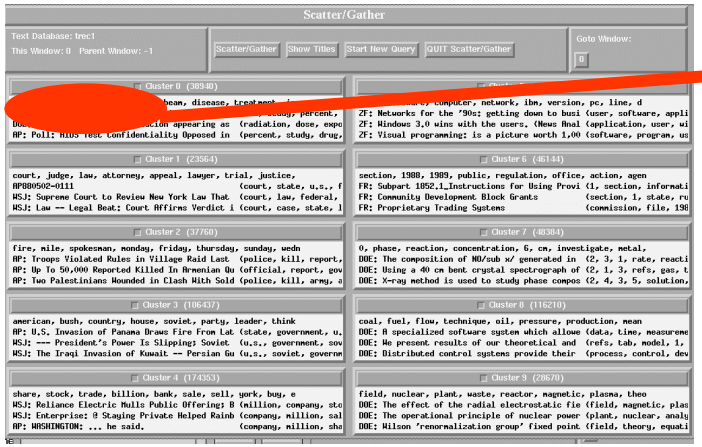
parc
Palo Alto Research Center

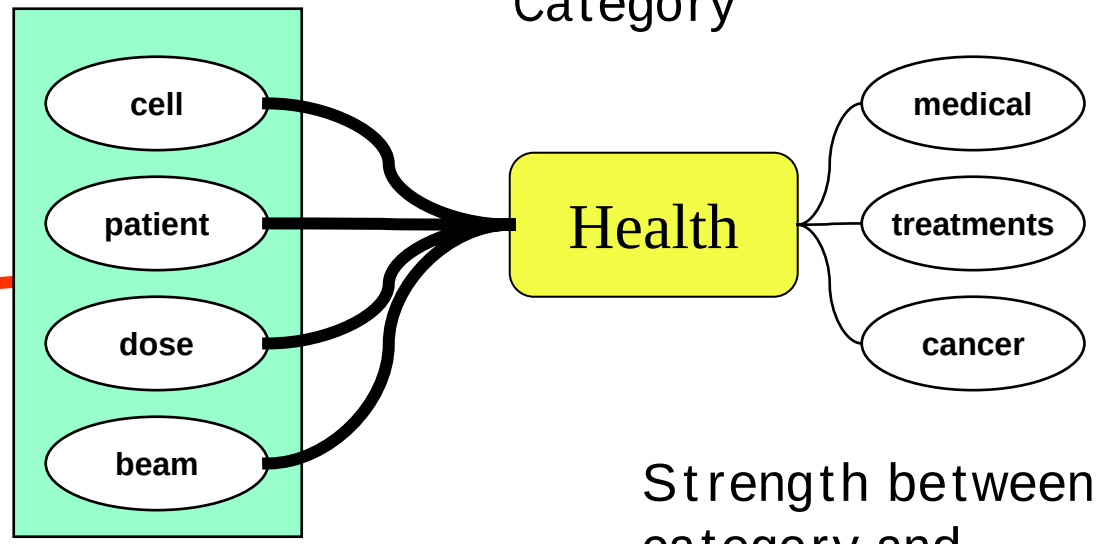
InfoCLASS

- Information Category Learning by Adaptation to Scent Stimuli
  - Rational Analysis Model of Human Category Learning (Anderson, 1991, 1991)
- Assumes
  - Topics are mental categories
  - Mental categories are probabilistic collections of concepts
  - Links, thumbnails, citations provide information scent cues that evoke mental concepts

parc
Palo Alto Research Center

Evoking Categories and Inferences



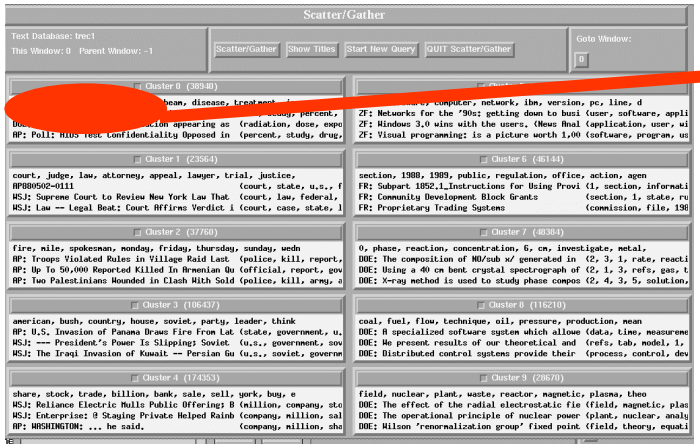


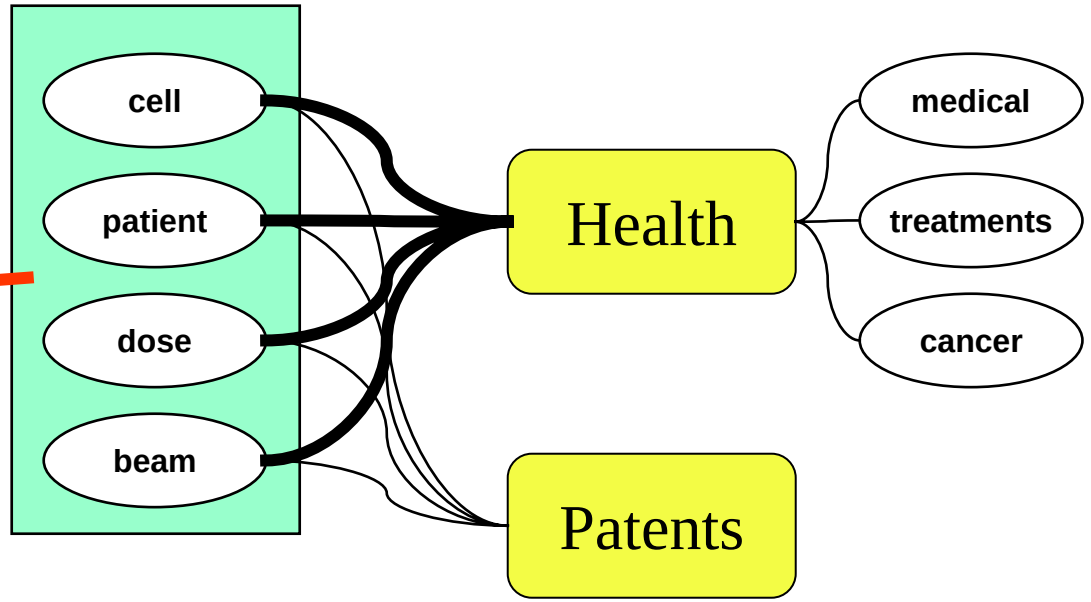
Strength between observed UI features (words) and category reflects log odds

Strength between category and inferred features reflects log odds

parc
Palo Alto Research Center

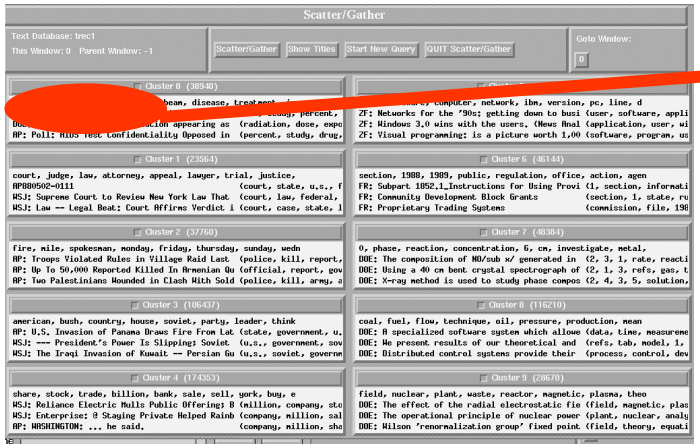
Cues correspond to existing mental topics: Choose most activated (highest odds)

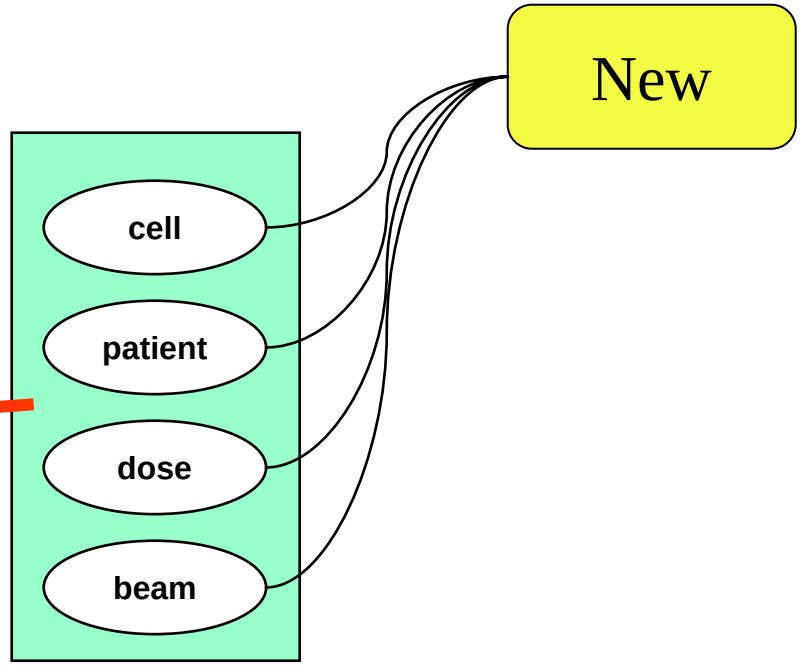




parc
Palo Alto Research Center

Otherwise create new category

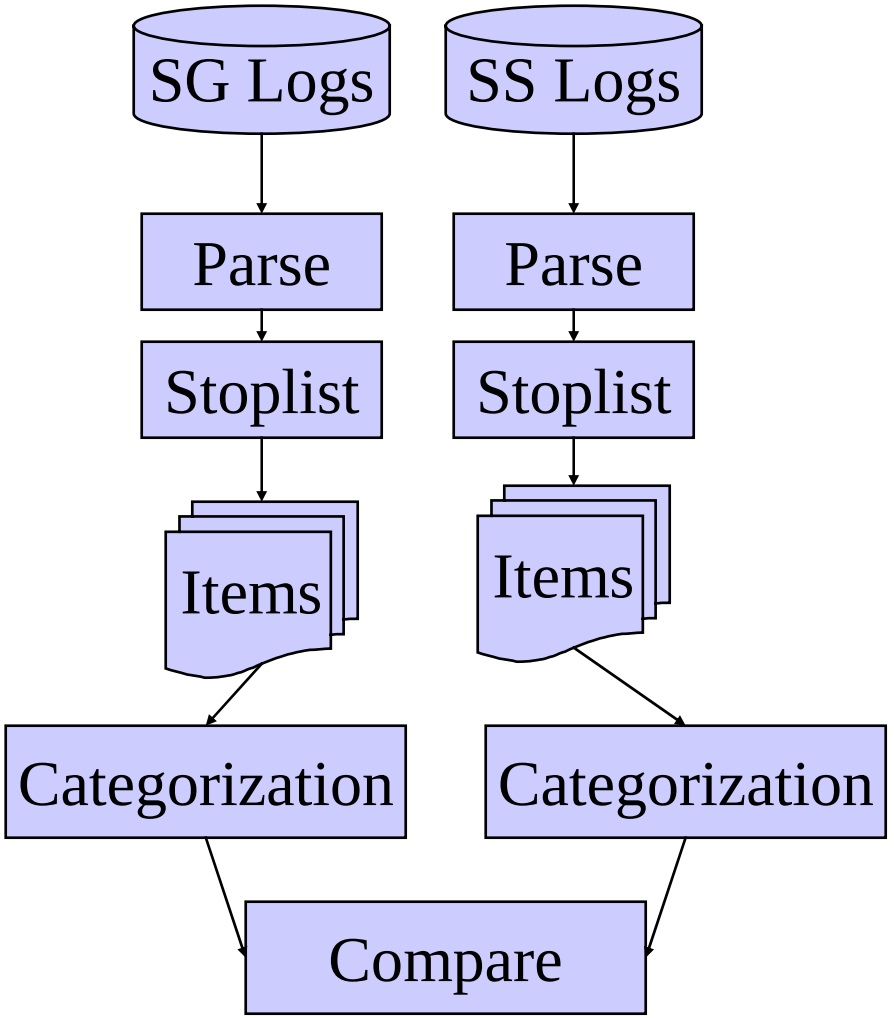




parc
Palo Alto Research Center

InfoCLASS Simulation

- Scatter/Gather (SG): expose to *cluster summaries* from user logs
- Standard Search (SS) Expose to *search result lists* from user logs
- Evaluate ADA (category coherence) for the two groups of users



parc
Palo Alto Research Center

Simulation Performance

	Items		Categories	
	SG	SS	SG	SS
1	330	464	59	28
2	470	767	149	39
3	526	1926	142	33
4	579	1481	101	28
5	770	1331	120	28
6	420	565	152	45
7	370	963	151	16
8	507	1486	124	14
	496	1123	125	29

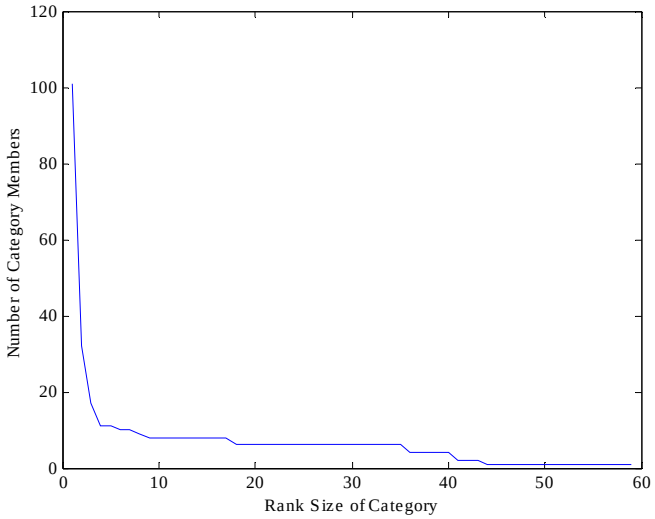
Scatter-Gather (SG) simulations develop more categories than Standard Search (SS).

Same trend as Pirolli et al (1996) comparison of SG vs SS users.

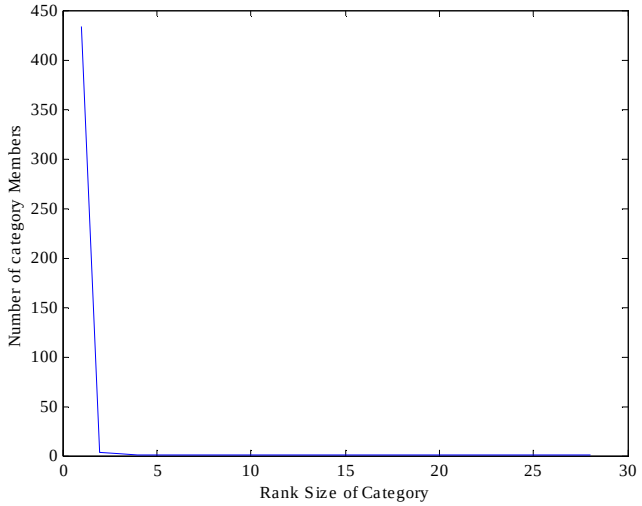
parc
Palo Alto Research Center

Category structure and coherence

Scatter/Gather User 1



Similarity Search User 1



	ADA
SG	.949
SS	1.292

x 10^{-6} bits

Scatter-Gather simulations exhibit greater category coherence (less divergence).

parc
Palo Alto Research Center

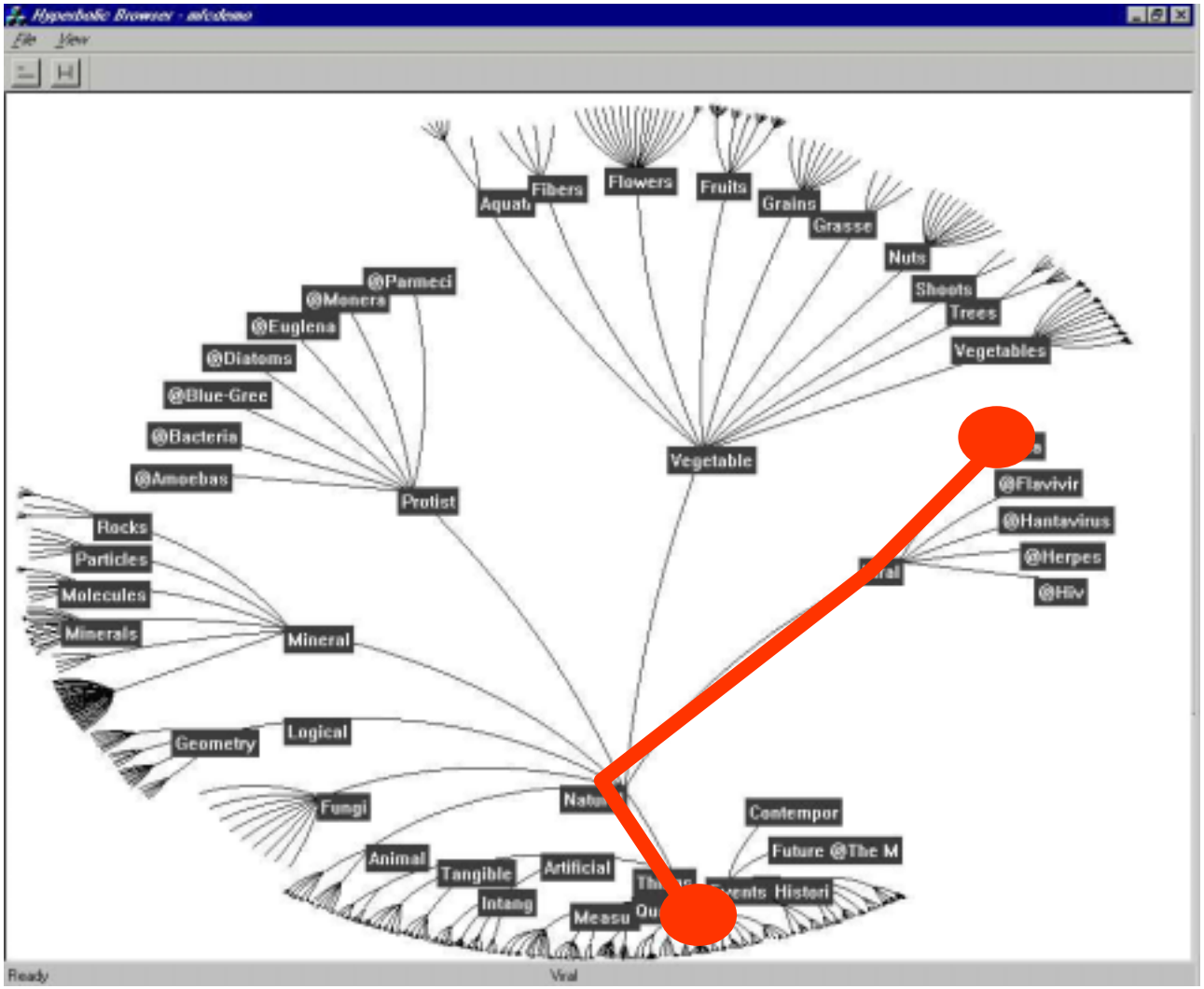
Summary

- Importance
  - Quality of information scent can effect qualitative changes in navigation costs (linear to exponential)
  - Browsers can differ in the how they communicate the topic structure of a collection of information
- SNIF-ACT model of navigation
  - Navigation & when to stop
- InfoClass
  - Model of formation of mental categories of topic structure
- Applications
  - New UIs (ScentTrails, Relevance Enhanced Thumbnails, …)
  - Web Site Evaluation tools (Bloodhound, Lumberjack)

BACKUP SLIDES

parc
Palo Alto Research Center

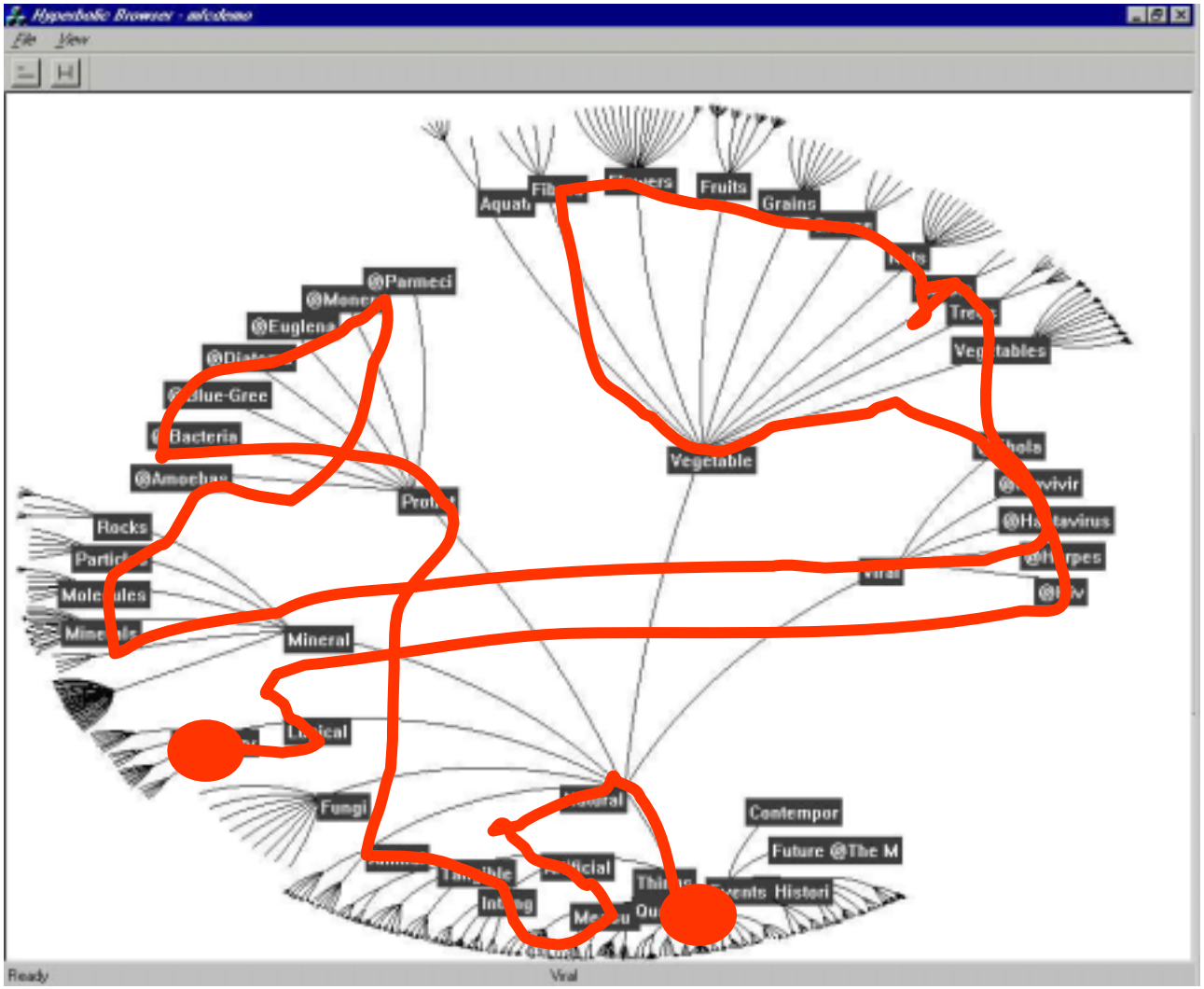
Eye Movements:
High Information Scent
(Pirolli, Card, & Van Der Wege, 2003, TOCHI)



parc
Palo Alto Research Center

Eye Movements:
Low Information Scent

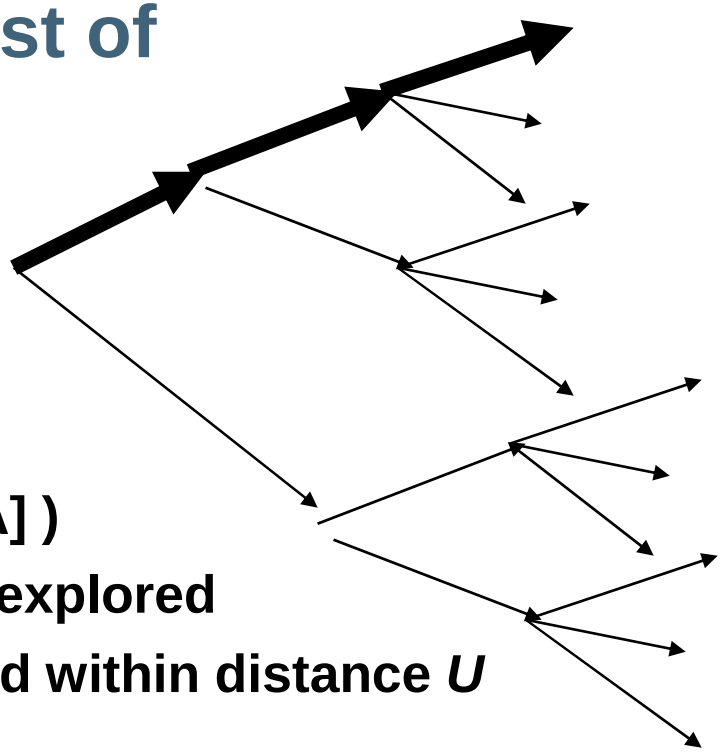
(Pirolli, Card, & Van Der Wege, 2003, TOCHI)



parc
Palo Alto Research Center

Information Scent and the Cost of Navigation

(based on Hogg & Huberman, 1987)



D = depth of search hierarchy

z = average branching factor

$(1-q)$ = prob. of false alarm (Pr[FA])

$\mu(q, z)$ = qz =average no. branches explored

$A(U, q, z)$ = average no. nodes explored within distance U

$= (1-\mu(q, z)^{U+1})/(1-\mu(q, z))$

$N(D, z, q)$ = average no. nodes examined before desired goal found

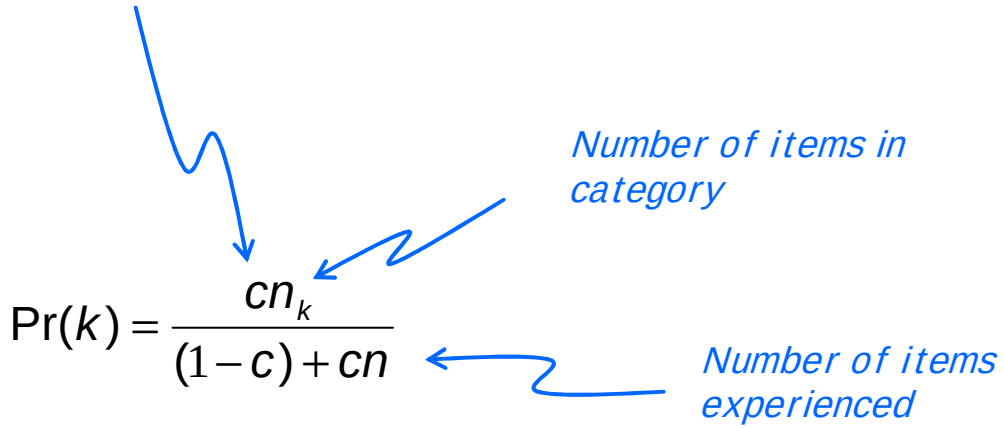
$$= \left[\frac{(z-1)q}{2}\right] \sum_{s=1}^{D-1} A(s-1, q, z)$$

parc
Palo Alto Research Center

Prototype Formation

- If there are no existing categories, then create a new category (k_{New}) and assign instance P to the category, otherwise
- Determine the prob that the instance comes from a new category, $\Pr(k_{New}|P)$, and compare that to the existing category with the highest probability of including the instance, $\Pr(k_{Max}|P)$
  - Assign P to k_{New} if $\Pr(k_{New}|P) > \Pr(k_{Max}|P)$, else
  - Assign P to k_{Max}

Coupling Parameter (c)



$$\Pr(k) = \frac{cn_k}{(1-c)+cn}$$

$$\Pr(k_{New}) = \frac{1-c}{(1-c)+cn}$$

parc
Palo Alto Research Center

Categorization Model

	w_1	w_2	....	w_N	
Item =	[2	0	1	0	0] Series of multinomial trials

	w_1	w_2		w_N	
Category =	[.5	.1	.4	.1	.1] Vector of probabilities

$$\Pr(p_1, p_2, \ldots p_n \mid \alpha_1, \alpha_2, \ldots \alpha_n) = \frac{1}{Z(\alpha_1, \alpha_2, \ldots \alpha)} \prod_i p_i^{\alpha_i - 1}$$ Dirichlet

$$\sum_i p_i = 1$$

Expected value $$\mathbf{E}[p_i] = \frac{\alpha_i}{\alpha_0}$$

$$\alpha_0 = \sum_i \alpha_i$$

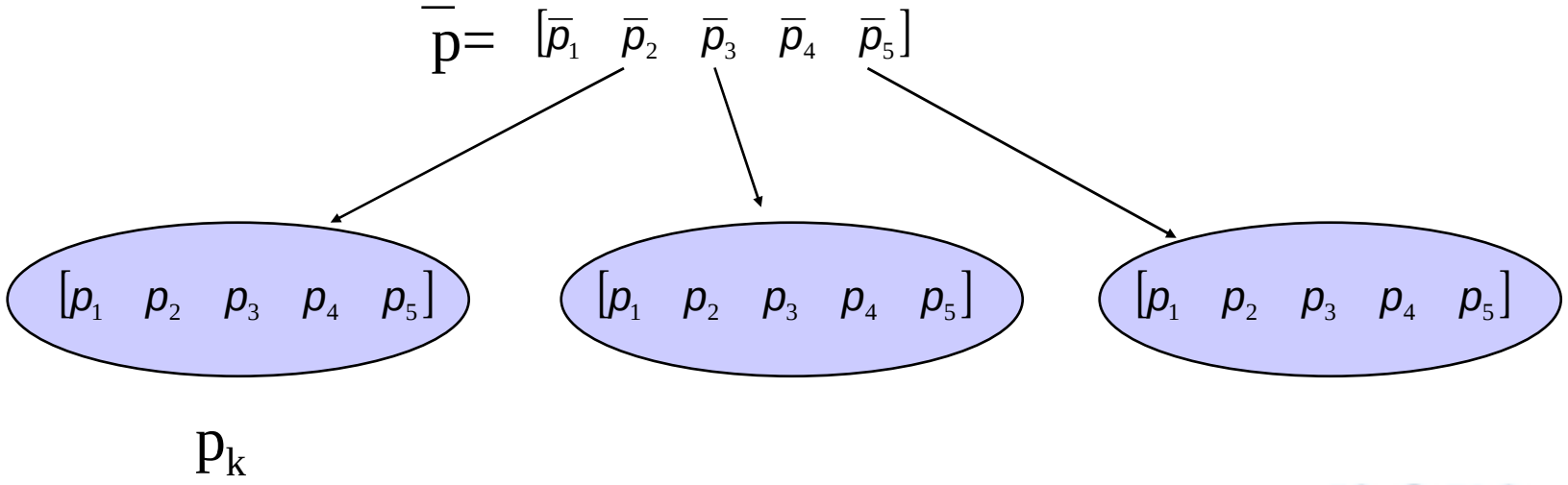
Noninformative prior $$\alpha_i = 1$$

Posterior distribution $$\mathbf{E}[p_i] = \frac{\alpha_i + n_i}{\alpha_0 + \sum_j n_j}$$

Average Divergence from Average Entropy

$$D(p \,||\, q) = \sum_x p(x) \log\left(\frac{p(x)}{q(w)}\right)$$

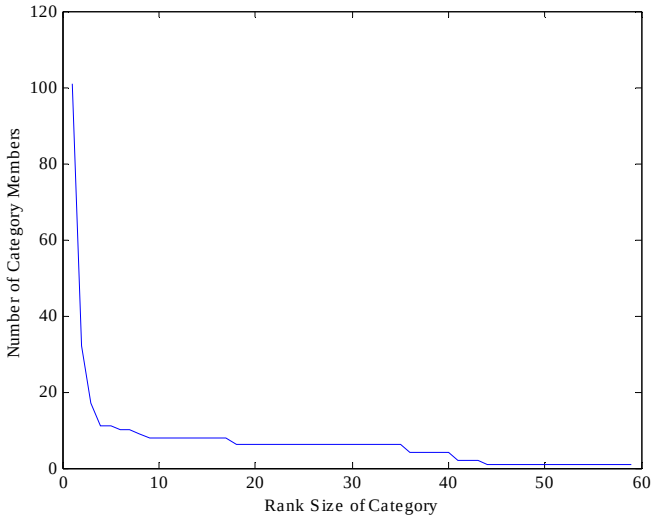
$$ADA = \frac{\sum_{k=1}^{K} D(\mathbf{p}_k \,||\, \bar{\mathbf{p}})}{K}$$



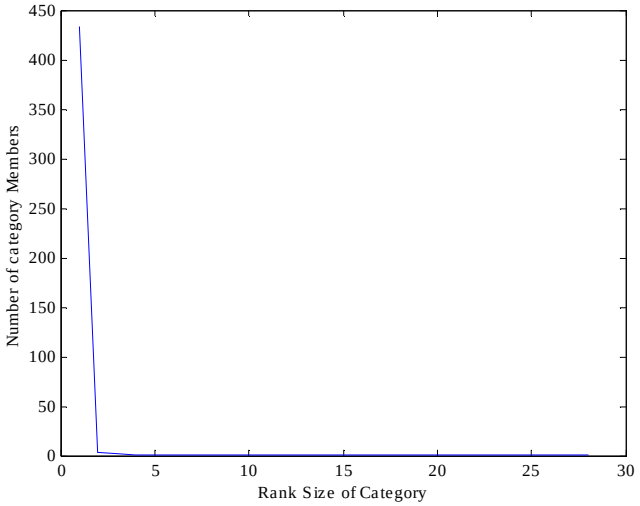
parc
Palo Alto Research Center

Category structure and coherence

Scatter/Gather User 1



Similarity Search User 1



	ADA
SG	.949
SS	1.292

x 10^{-6} bits

parc
Palo Alto Research Center