

Focused Crawling and Collection Synthesis

Donna Bergmark

Cornell Information Systems

bergmark@cs.cornell.edu

Outline

- Crawlers
- Collection Synthesis
- Focused Crawling
- Some Results
- Student Project (Fall 2002)

Resource Discovery

- Finding info on the Web
 - Surfing (random strategy; goal is serendipity)
 - Searching (inverted indices; specific info)
 - Crawling (follow links; “all” the info)
- Uses for crawling
 - Find stuff
 - Gather stuff
 - Check stuff

Definition

Spider = robot = crawler

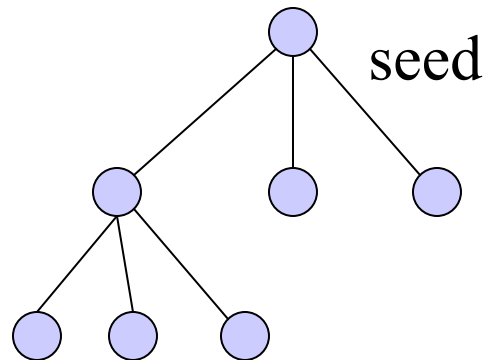
Crawlers are computer programs that roam the Web with the goal of automating specific tasks related to the Web.

Crawlers and internet history

- 1991: HTTP
- 1992: 26 servers
- 1993: 60+ servers; self-register; archie
- 1994 (early) – first crawlers
- 1996 – search engines abound
- 1998 – focused crawling
- 1999 – web graph studies
- 2002 – use for digital libraries

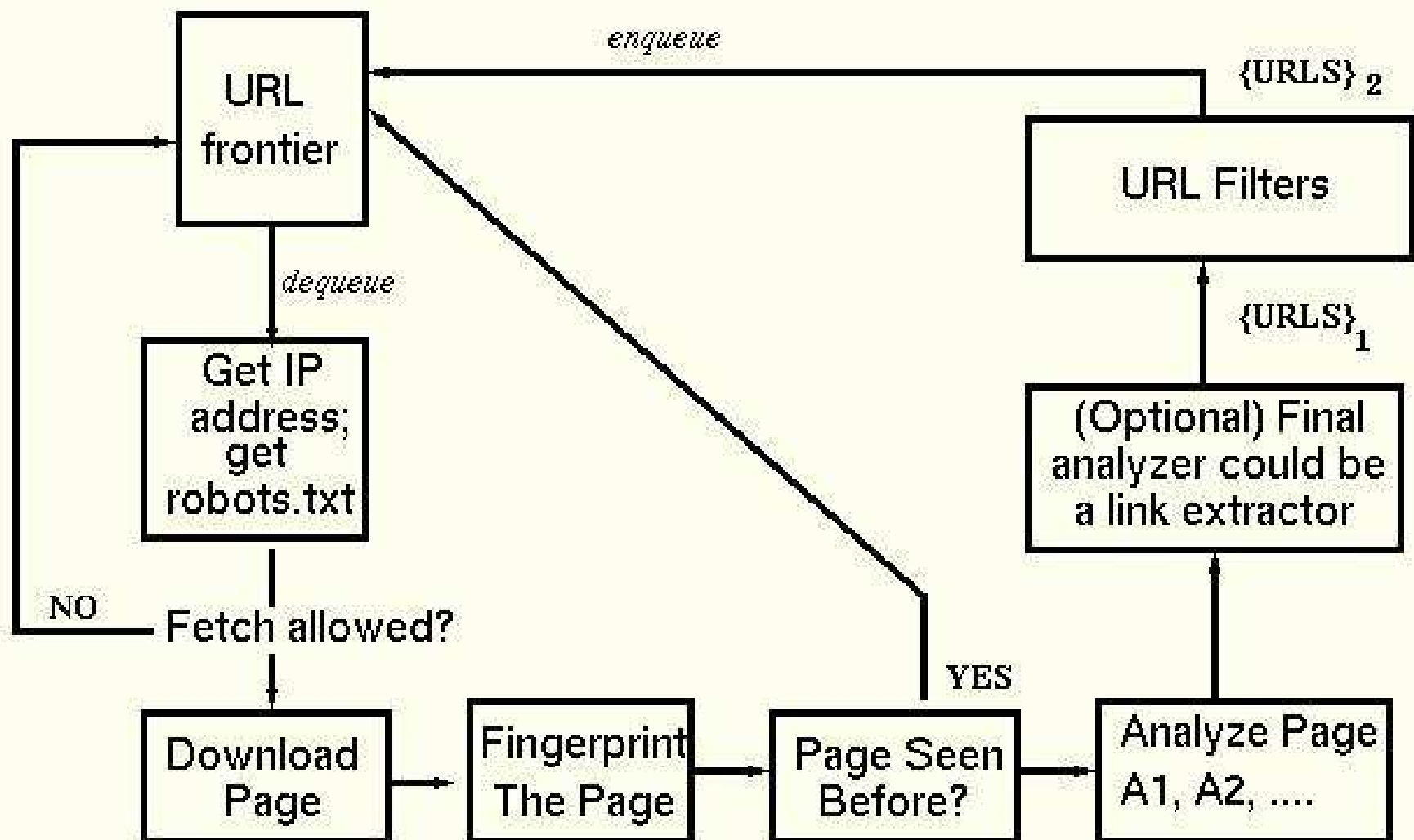
Crawling and Crawlers

- Web overlays the internet
- A crawl overlays the web



Crawler Issues

- The web is so big
- Visit order
- The URL itself
- Politeness
- Robot Traps
- The hidden web
- System Considerations



Mercator Features

- Written in Java
- One file configures a crawl
- Can add your own code
 - Extend one or more of M's base classes
 - Add totally new classes called by your own
- Industrial-strength crawler:
 - uses its own DNS and java.net package

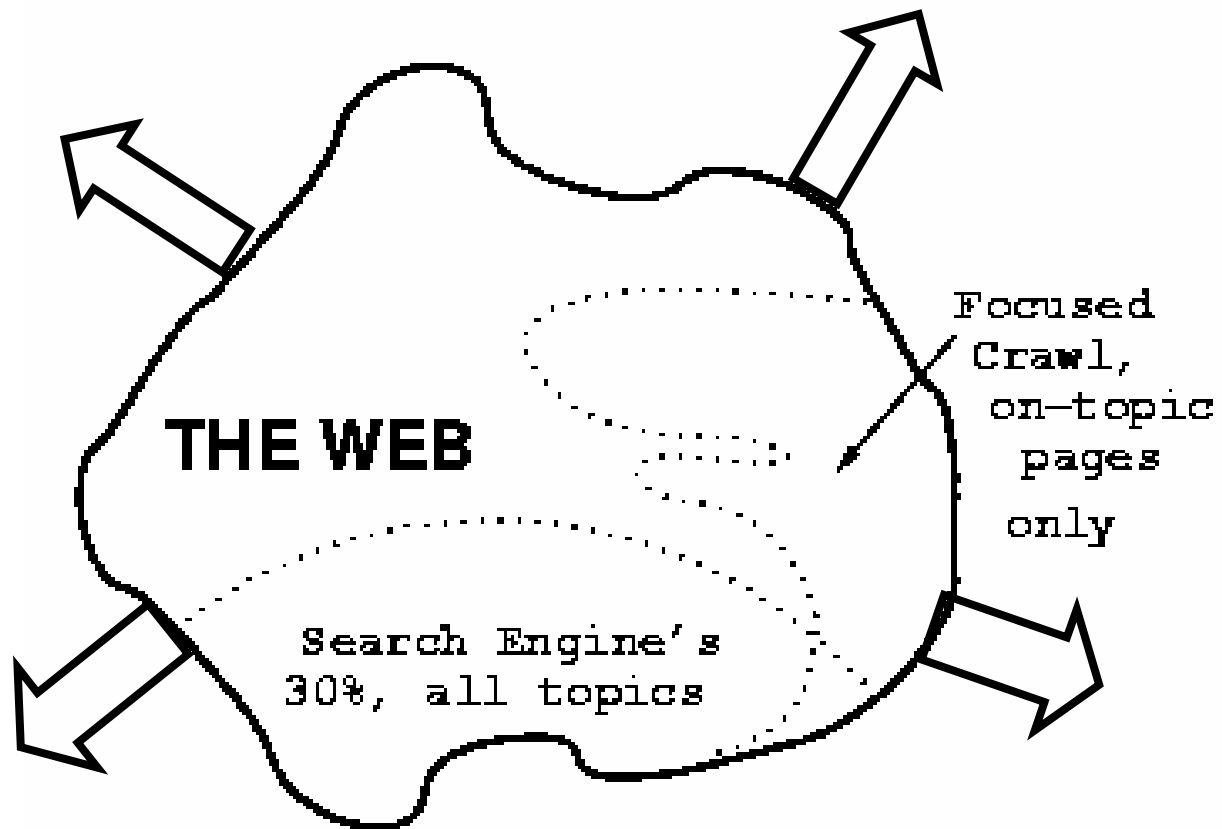
Collection Synthesis

- The NSDL
 - National Scientific Digital Library
 - Educational materials for K-thru-grade
 - A collection of digital collections
- Collection (automatically derived)
 - 20-50 items on a topic, represented by their URLs, expository in nature, precision trumps recall.
- Collection description (automatically derived)

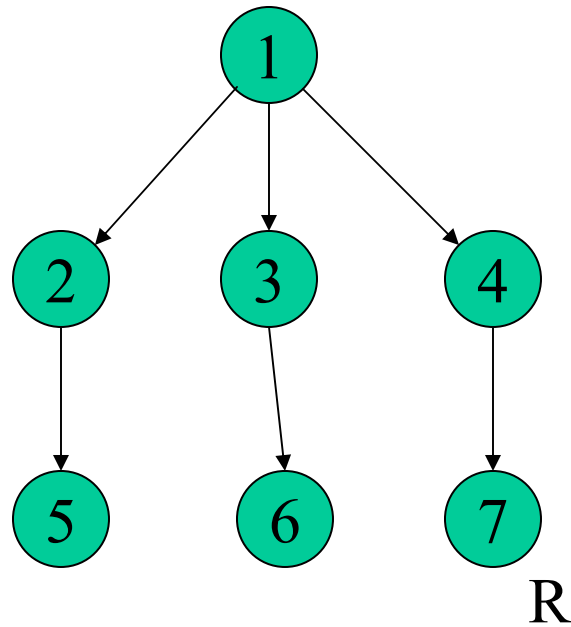
Crawler is the Key

- A general search engine is good for precise results, few in number
- A search engine must cover all topics, not just scientific
- For automatic collection assembly, a powerful Web crawler is needed
- A **focused crawler** is the key

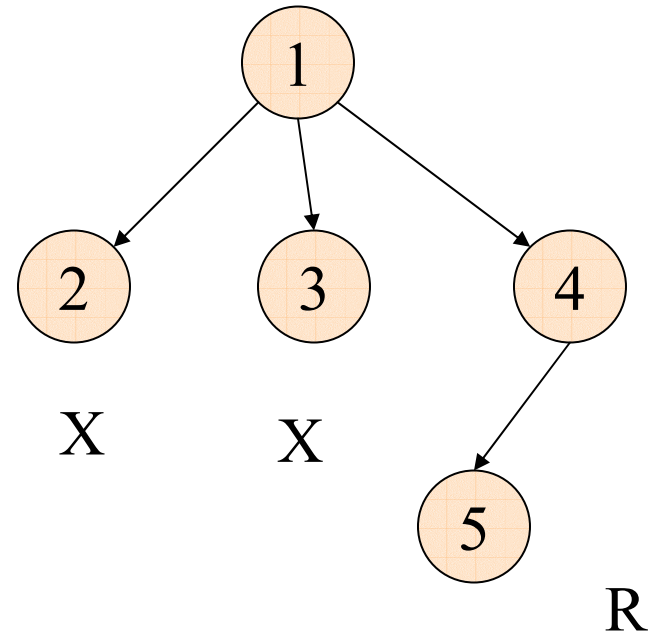
Focused Crawling



Focused Crawling



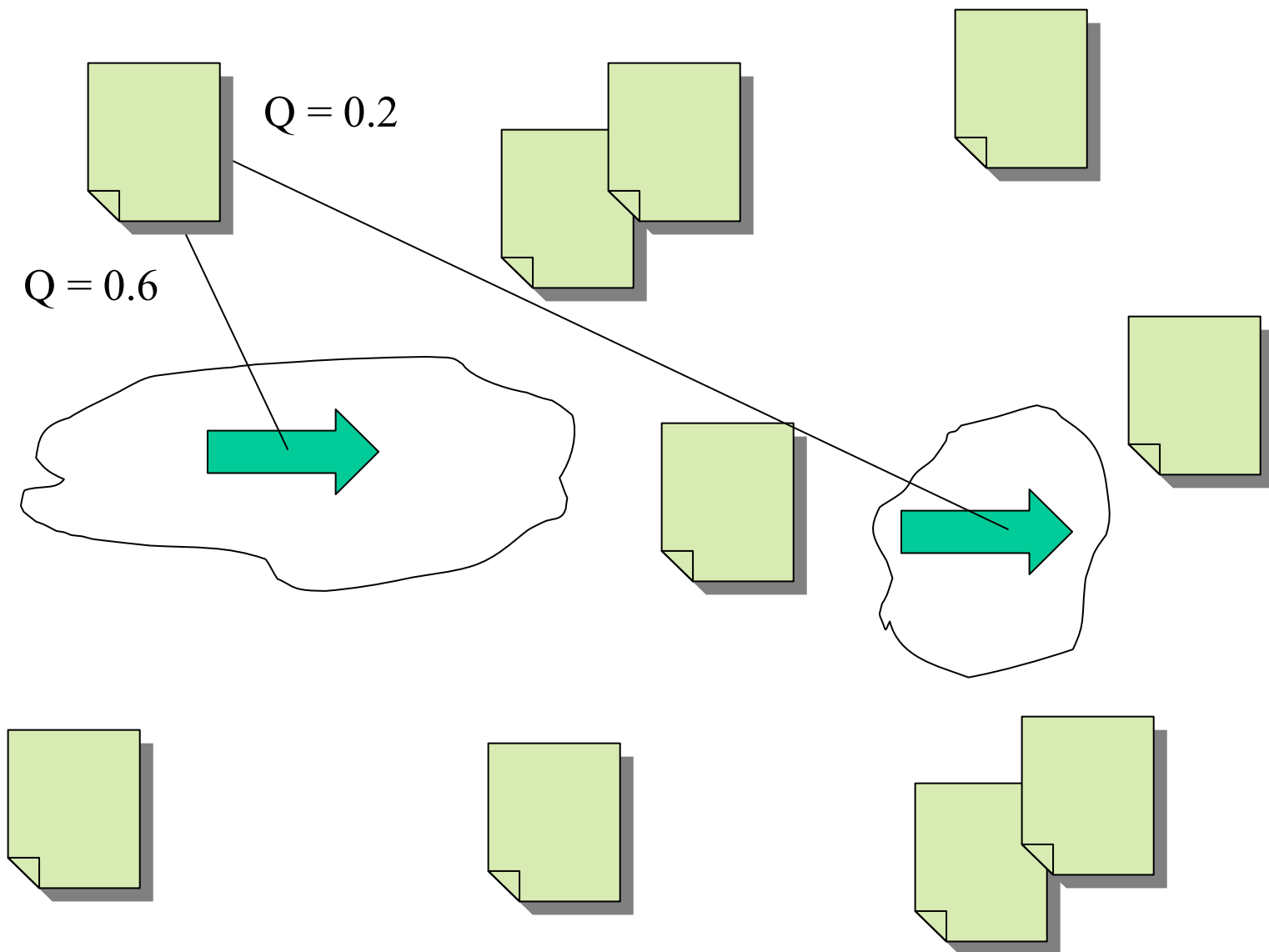
1 Breadth-first crawl



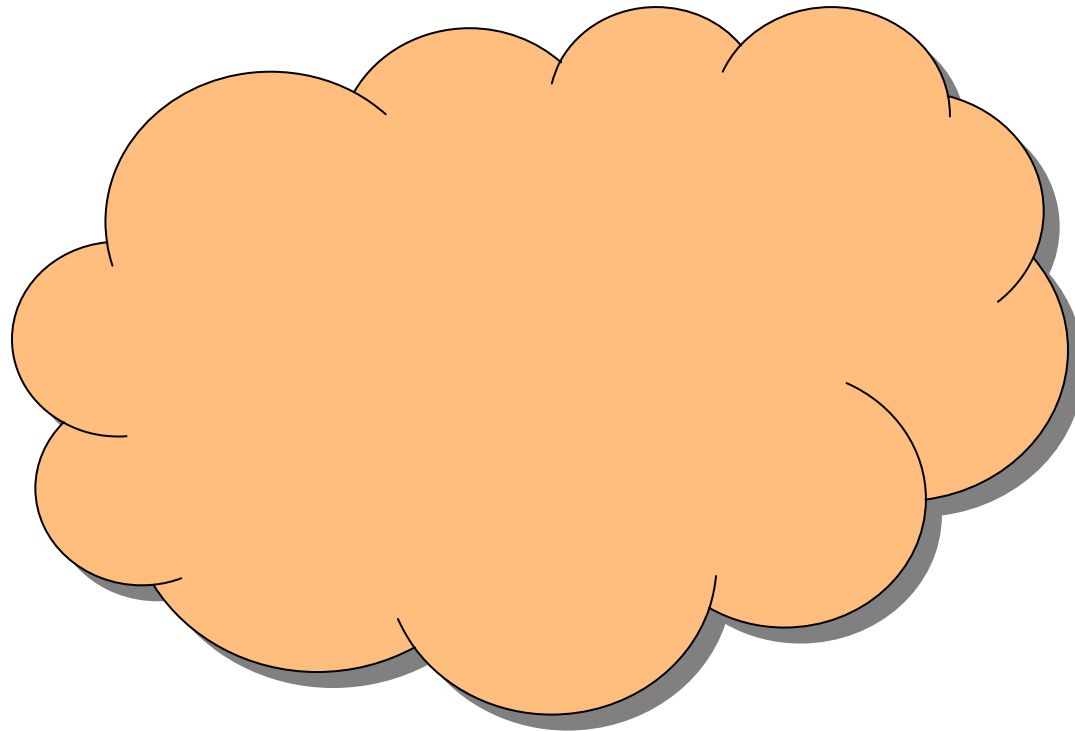
Focused crawl

Collections and Clusters

- Traditional – document universe is divided into clusters, or collections
- Each collection represented by its **centroid**
- Web – size of document universe is infinite
- Agglomerative clustering is used instead
- Two aspects:
 - Collection descriptor
 - Rule for when items belong to that Collection

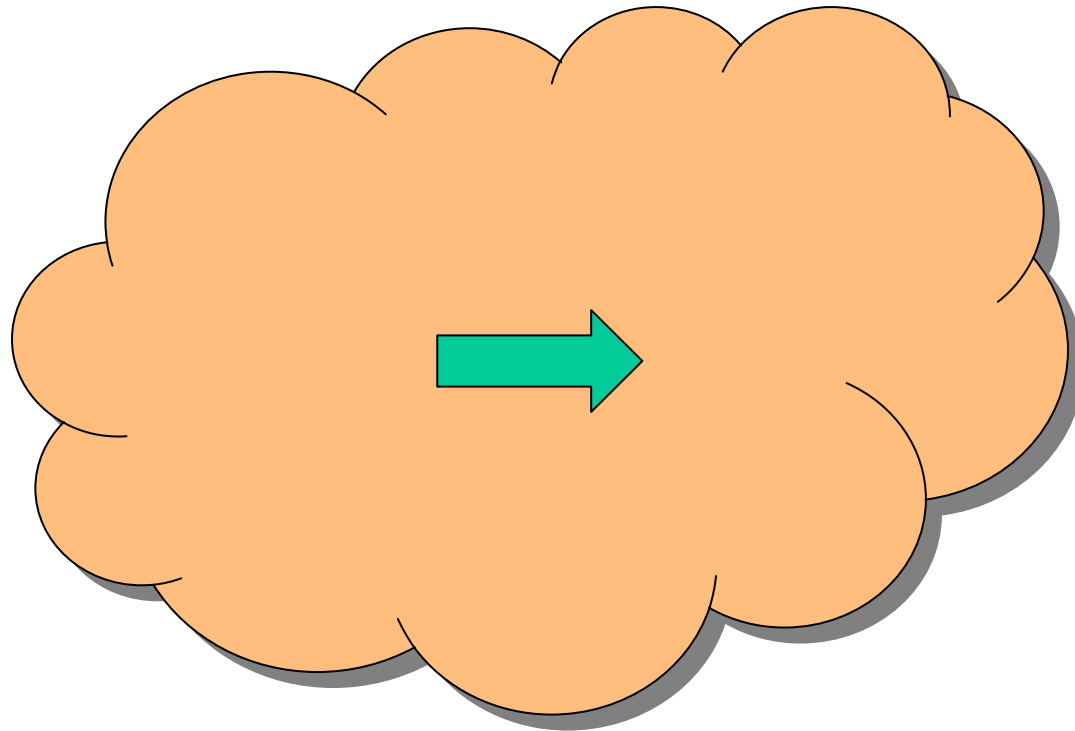


The Setup



A virtual collection of items about Chebyshev Polynomials

Adding a Centroid



An empty collection of items about Chebyshev Polynomials

Document Vector Space

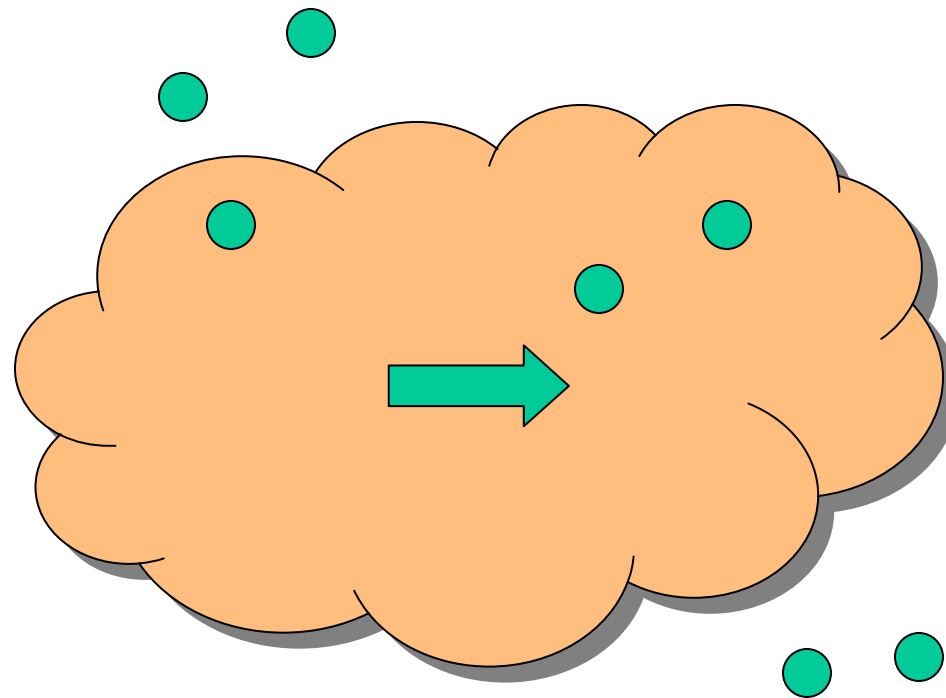
- Classic information retrieval technique
- Each word is a dimension in N-space
- Each document is a vector in N-space

Example: $\langle 0, 0.003, 0, 0, .01, .984, 0, .001 \rangle$

- Normalize the weights

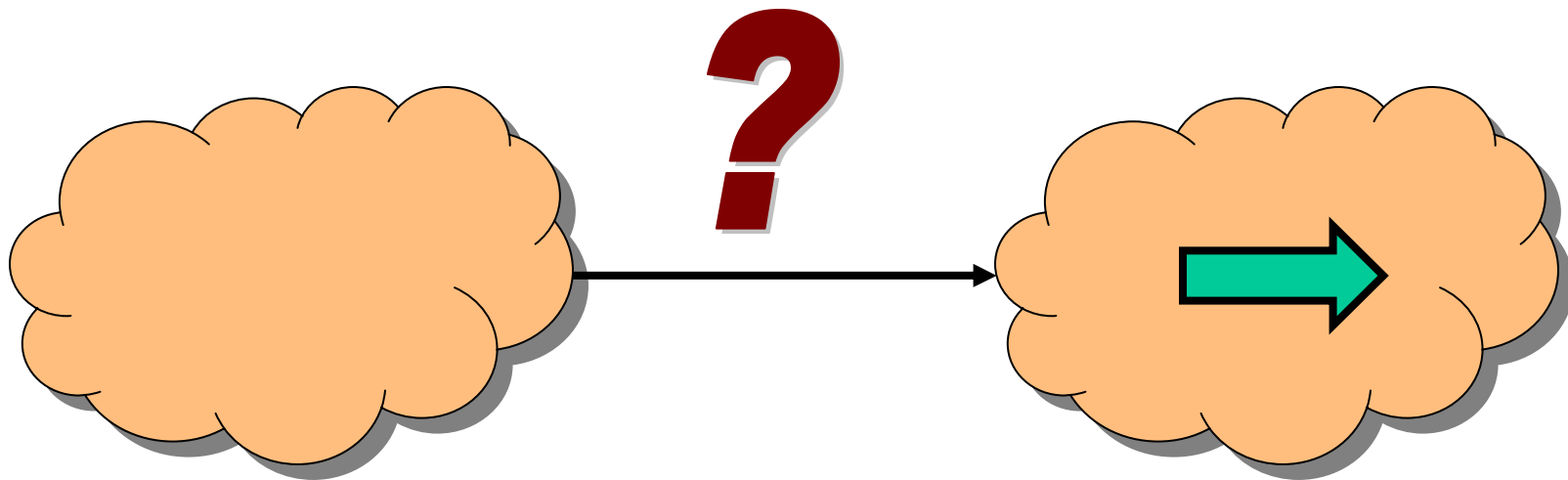
Both the “centroid” and the downloaded document are term vectors

Agglomerate



A collection with 3 items about Ch. Polys.

Where does the Centroid come from?



“Chebyshev
Polynomials”

A really good centroid for
a collection about C.P.’s

Building a Centroid

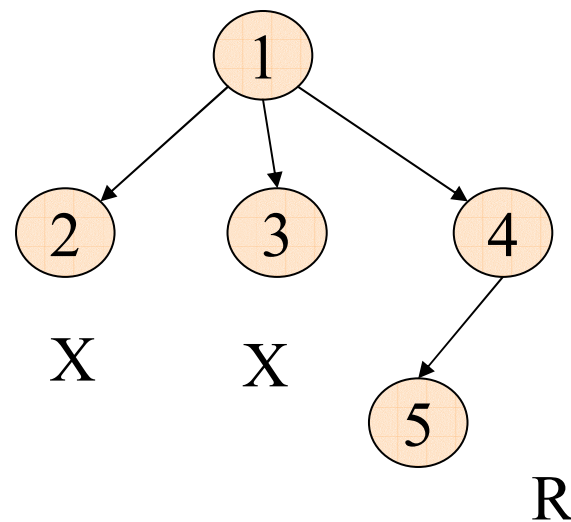
1. Google(“Chebyshev Polynomials”) \rightarrow url1, url2, ...
2. Let H be a hash (k,v) where k=word, value=freq
3. For each url in {url1, url2,...} do
 - D \leftarrow download(url)
 - V \leftarrow term vector(d)
 - For each term t in V do
 - If t not in H add it with value 0
 - H(t) ++
4. Compute tf-idf weights. C \leftarrow top 20 terms (by weight).

Dictionary

- Given centroids $C_1, C_2, C_3 \dots$
- Dictionary is $C_1 + C_2 + C_3 \dots$
 - Terms are union of terms in C_i
 - Document Frequency is how many C 's have t
 - Term IDF is based on Berkeley's DocFreqs
- Dictionary is 300-500 terms

Focused Crawling

- Recall the cartoon for a focused crawl:



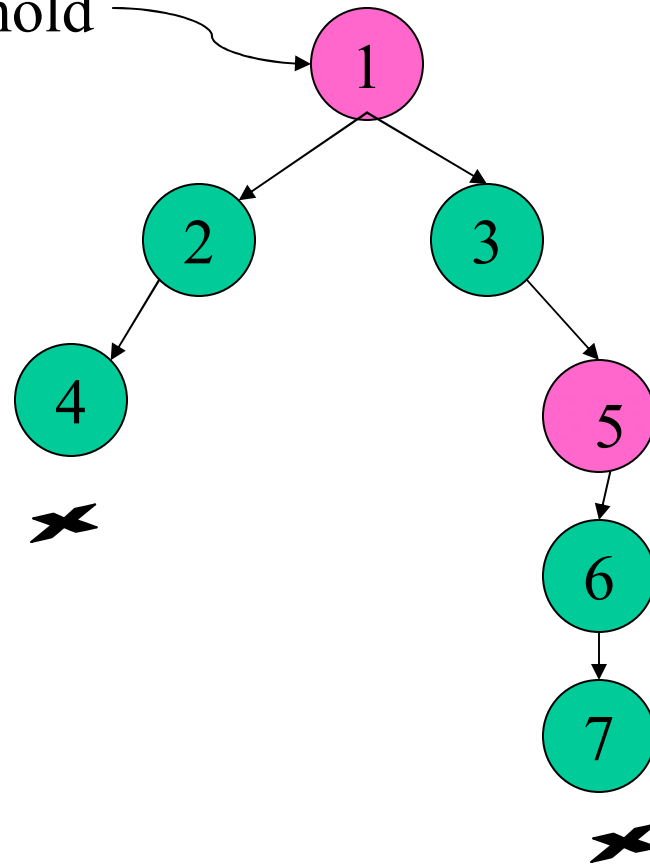
- A simple way to do it is with 2 “knobs”

Focusing the Crawl

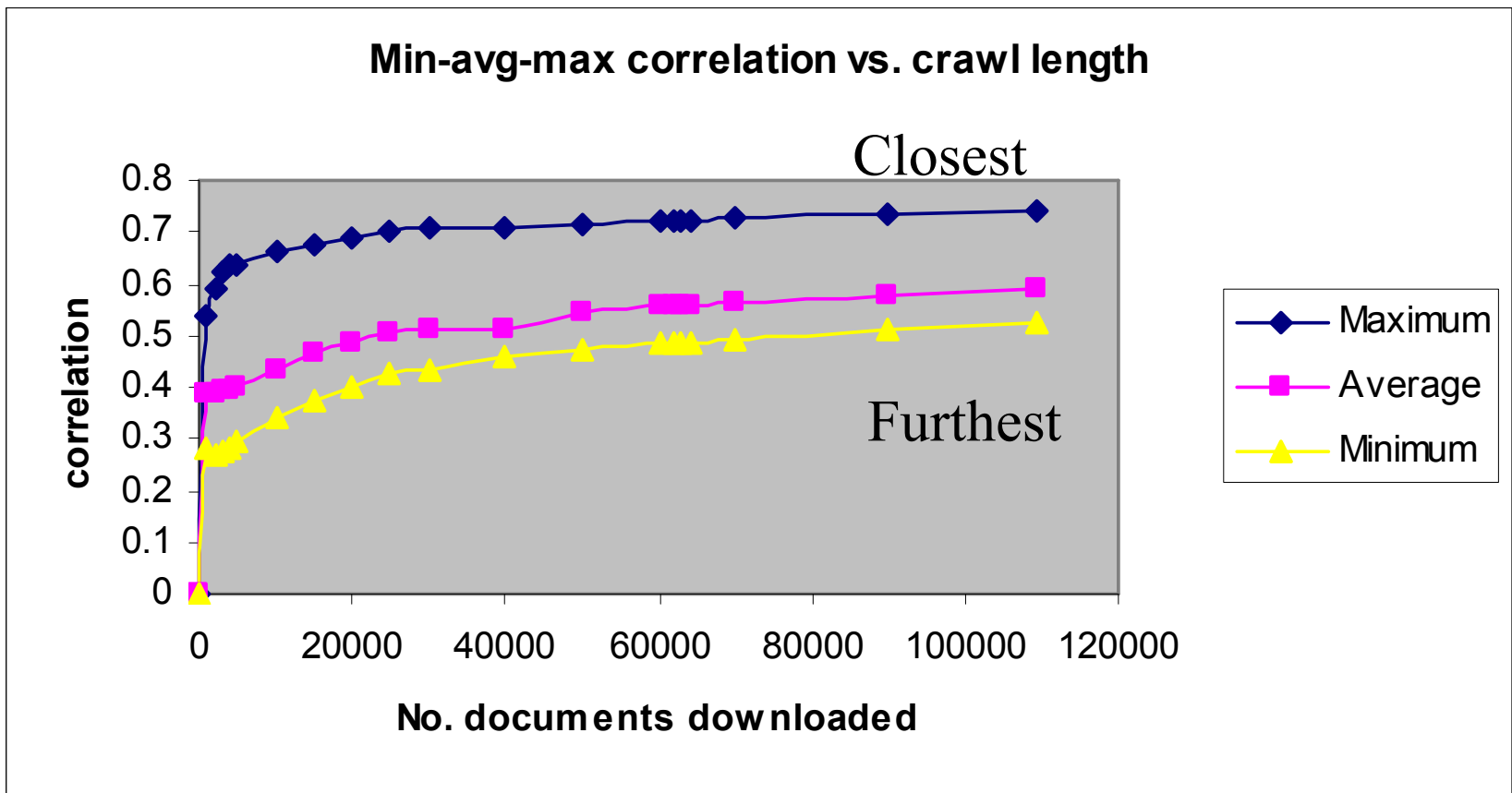
- **Threshold**: page is on-topic if correlation to the closest centroid is above this value
- **Cutoff**: follow links from pages whose “distance” from closest on-topic ancestor is less than this value

Illustration

Corr \geq threshold

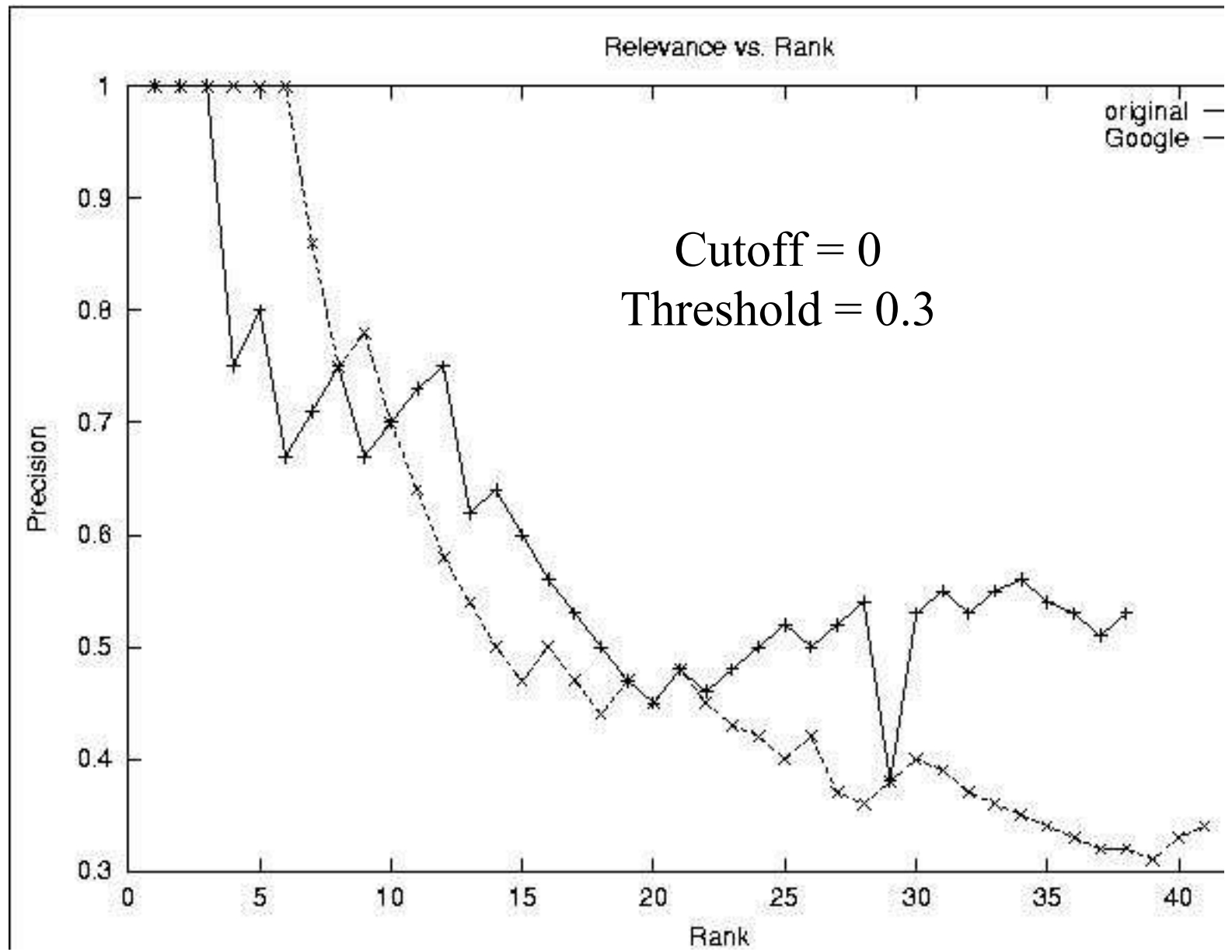


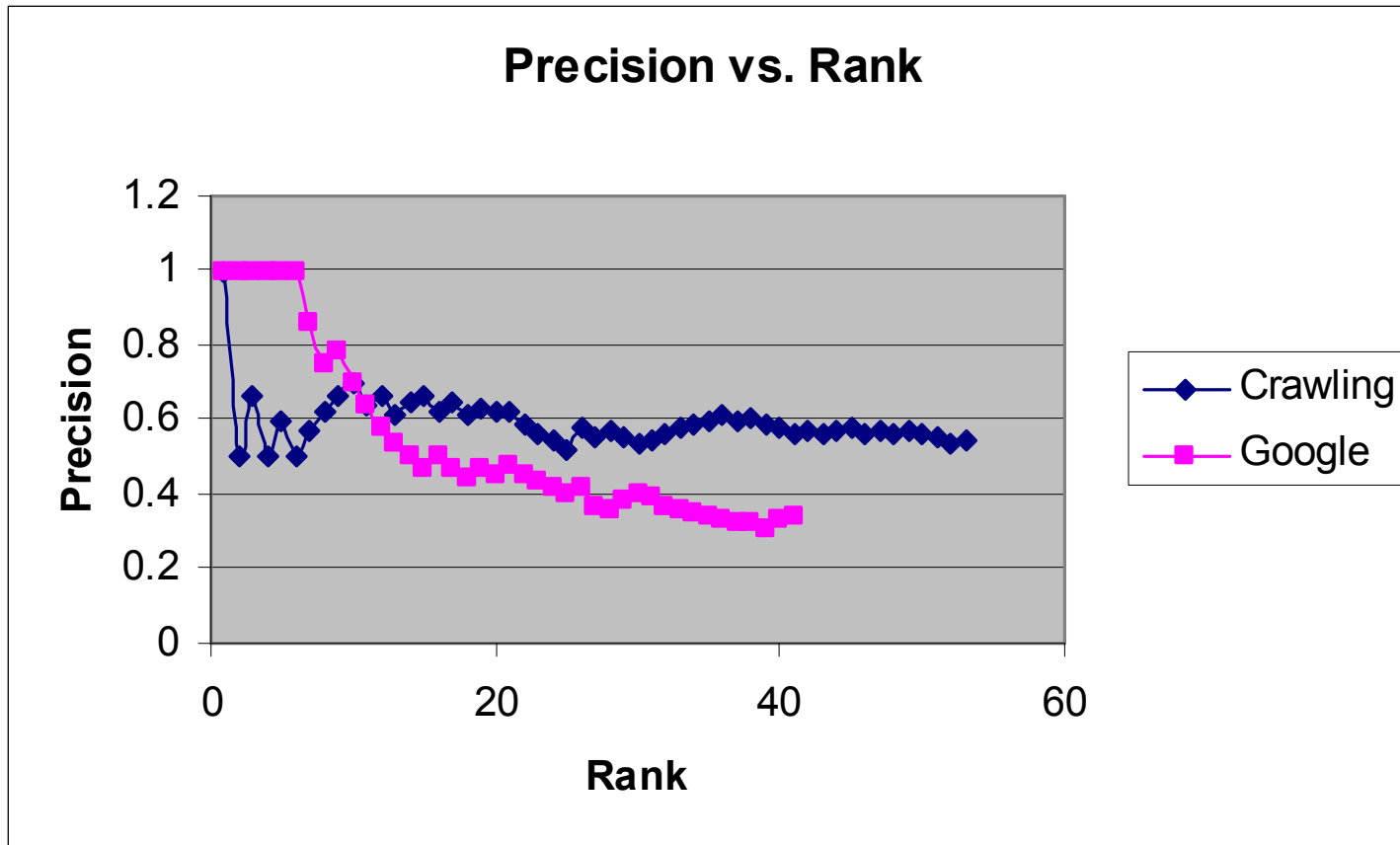
Cutoff = 1



Collection “Evaluation”

- Assume higher correlations are good
- With human relevance assessments, one can also compute a “precision” curve
- Precision $P(n)$ after considering the n most highly ranked items is number of relevant, divided by n .





Tunneling with Cutoff

- Nugget – dud – dud... – dud – nugget

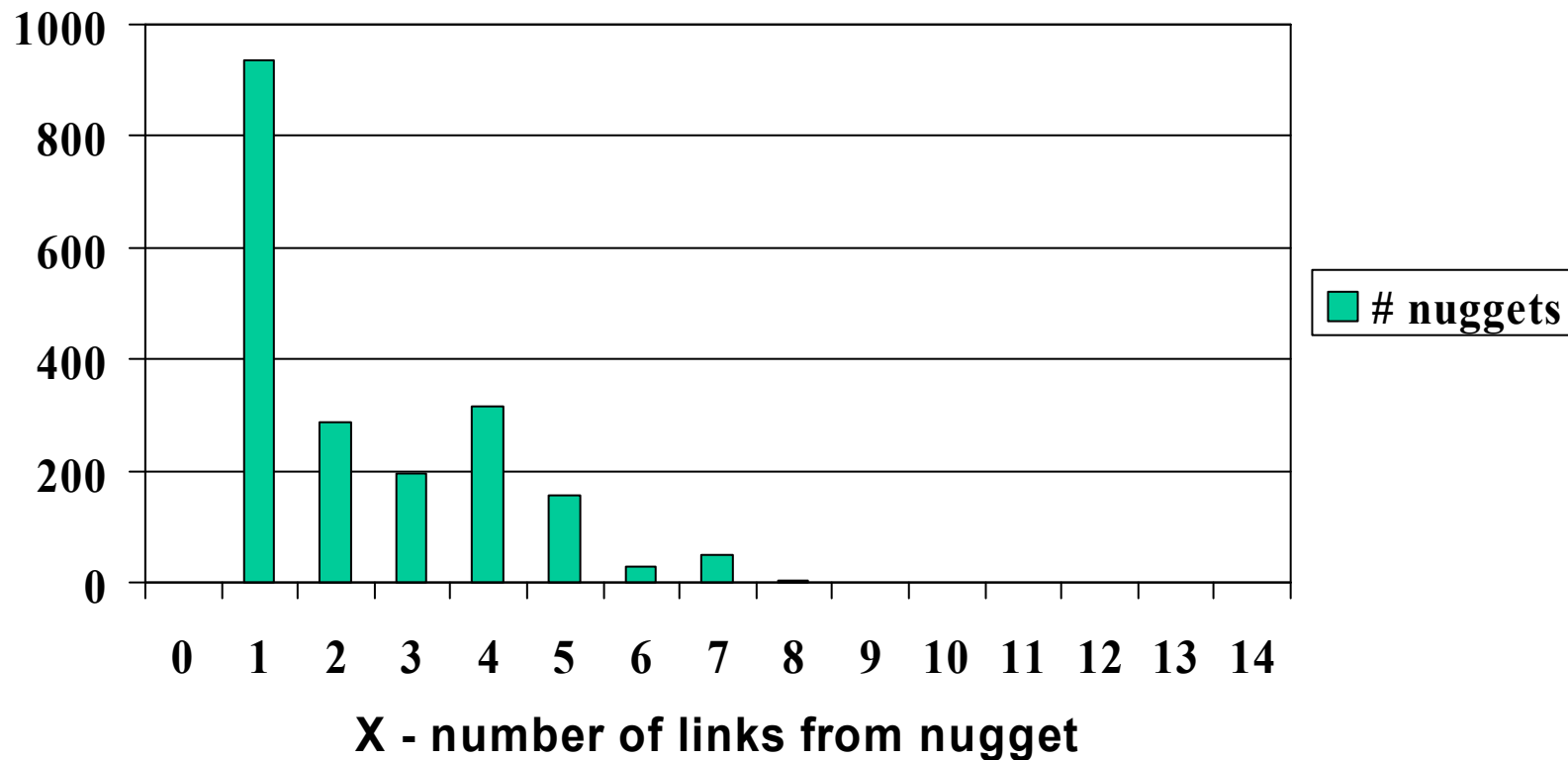
Notation: $0 - X - X \dots - X - 0$

- Fixed cutoff: $0 - X_1 - X_2 - \dots - X_c$
- Adaptive cutoff: $0 - X_1 - X_2 - \dots - X?$

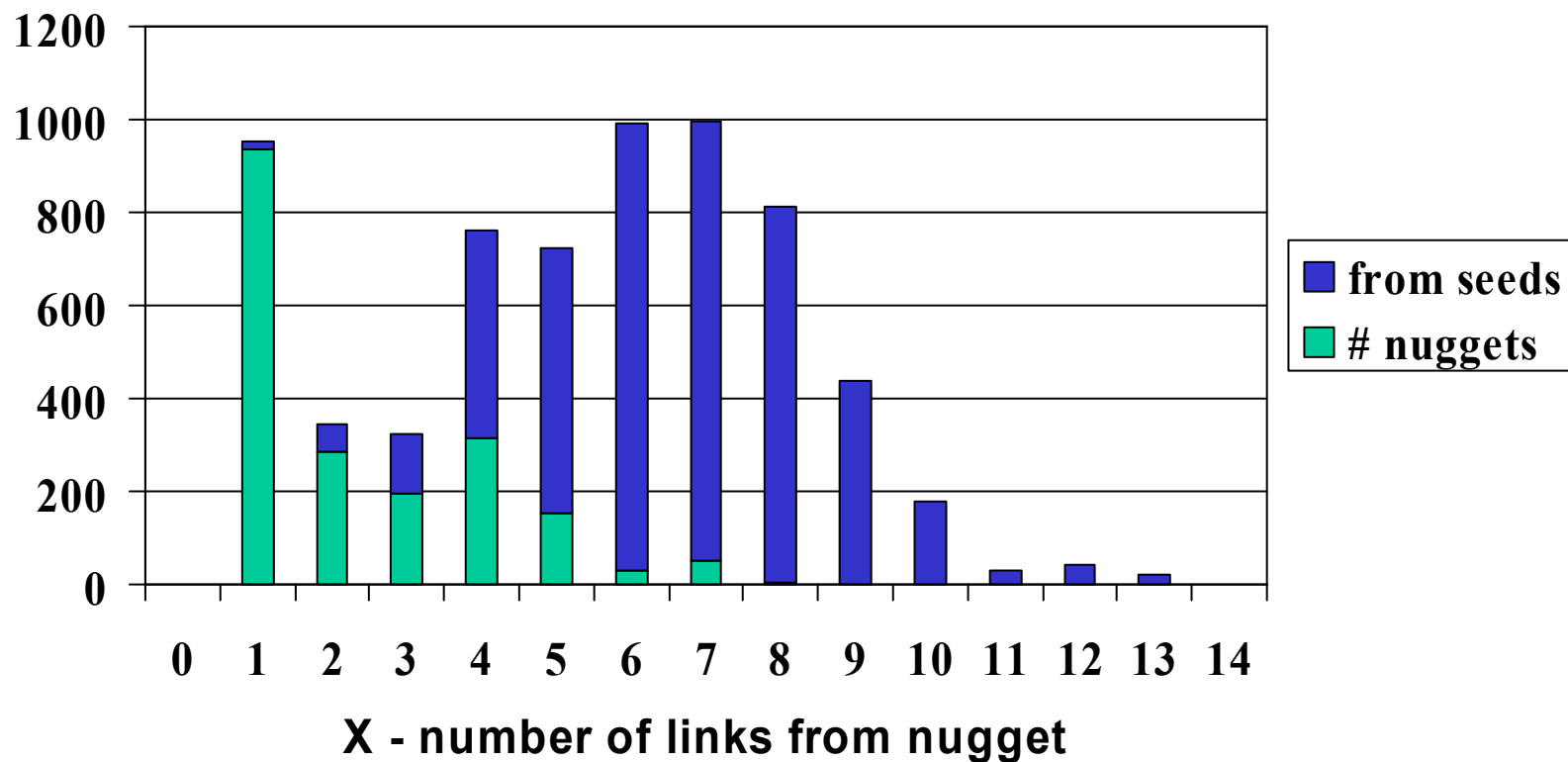
Statistics Collected

- 500,000 documents
- Number of seeds: 4
- Path data for all but seeds
- 6620 completed paths (0-x...x-0)
- 100,000s incomplete paths (0-x...x..)

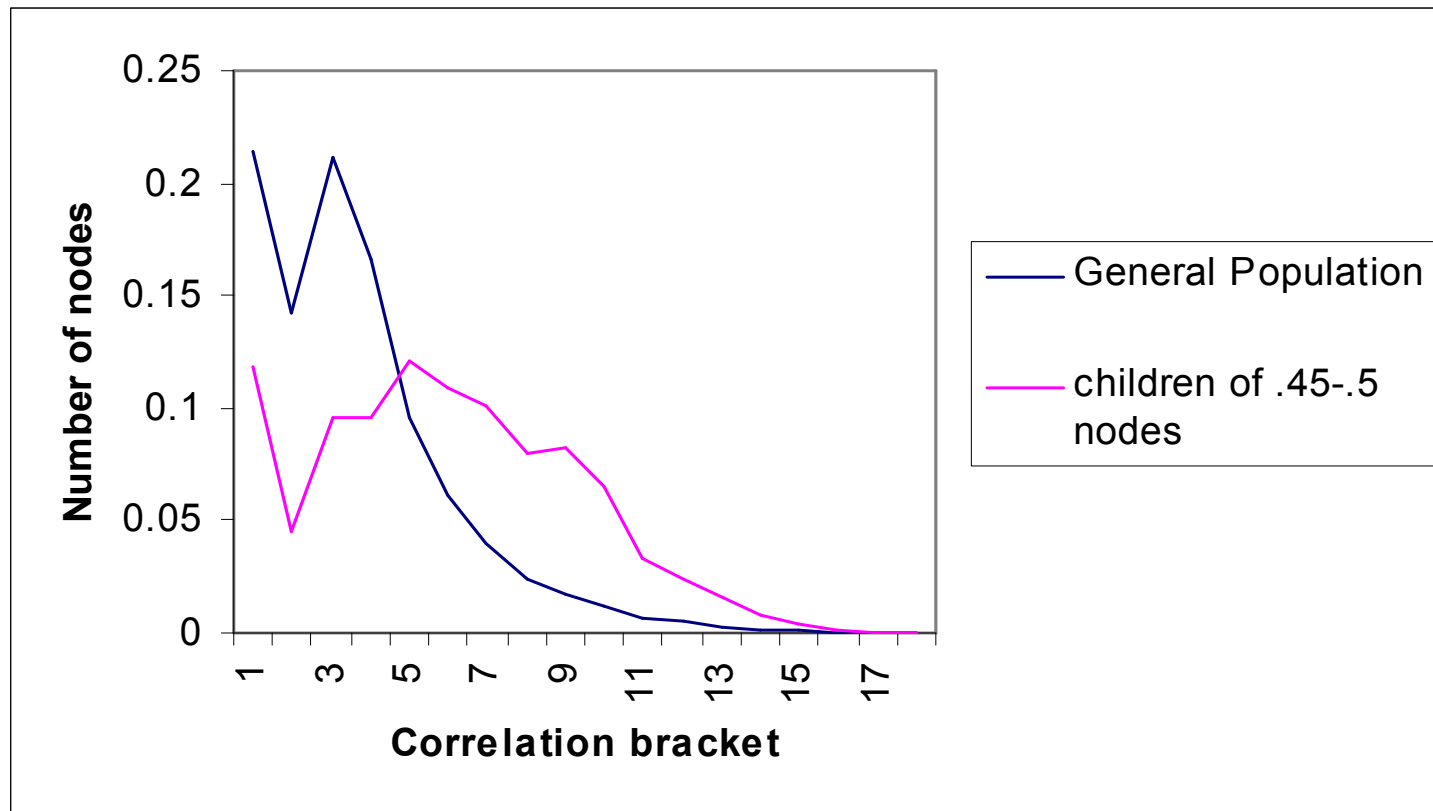
Nuggets that are x steps from a nugget



Nuggets that are x steps from a seed and/or a nugget



Better parents have better children.



Using the Empirical Observations

- Use path history
- Page quality = cosine correlation
- Current distance should increase exponentially as you get away from quality nodes

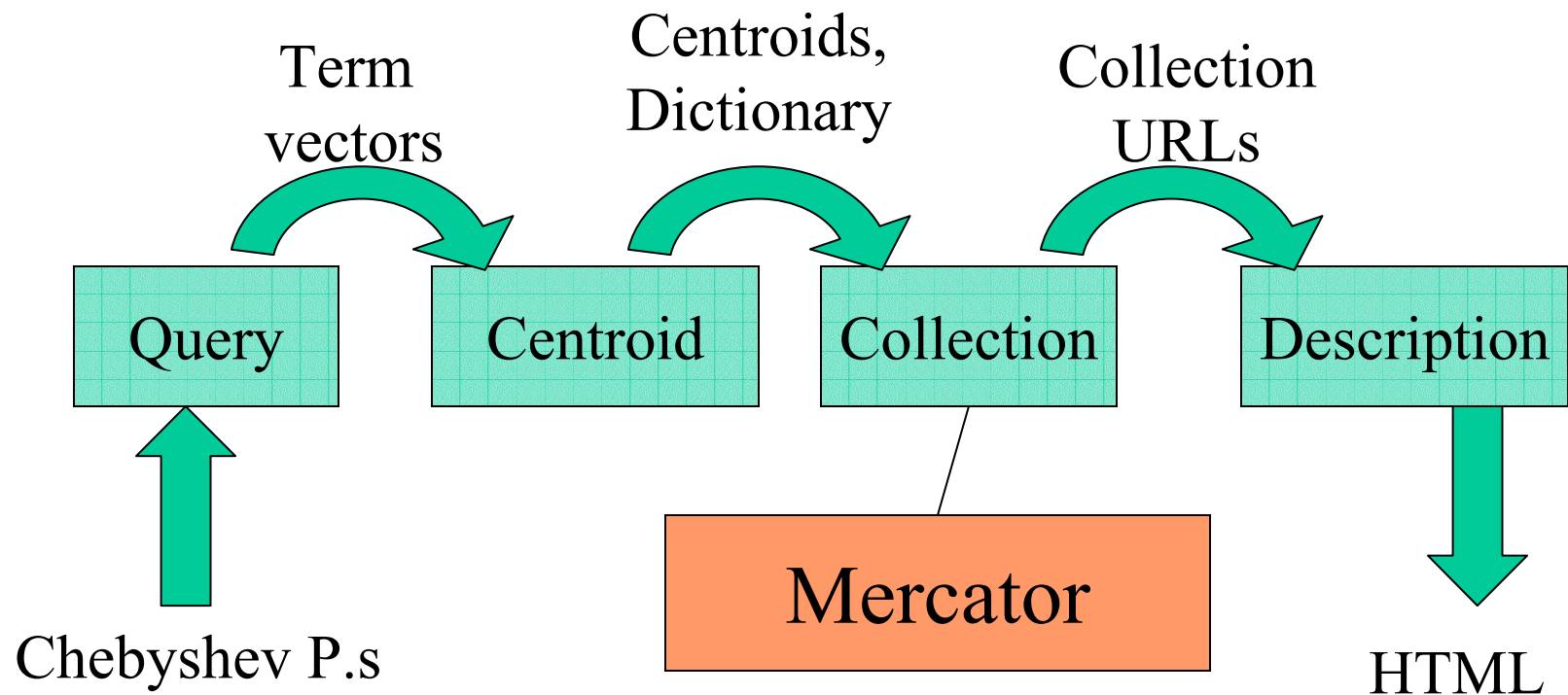
Distance = 0 if this is a nugget, otherwise:

1 or (1-corr) exp (2 x parent's distance / cutoff)

Results

- Details in our ECDL 2002 paper
- Smaller frontier → more docs/second
- More documents downloaded in same time
- Higher-scoring documents were downloaded
- Cutoff of 20 averaged 7 steps at the cutoff

Fall 2002 Student Project



Conclusion

- We covered crawling – history, technology, deployment
- Focused crawling with tunneling
- Adaptive cutoff with tunneling
- We have a good experimental setup for exploring automatic collection synthesis

<http://mercator.comm.nsdlib.org>