

Focused Crawling: a new approach to topic-specific Web resource discovery

Chakrabarti, Berg, and Dom
presented by Chad Hogg
2005-09-13

Why focused crawlers?

- ◆ Generic crawlers only index 30-40% of the web
- ◆ Crawl data is typically 3-4 weeks old
- ◆ Focused crawler runs in a few days on commodity hardware
- ◆ Many focused crawlers may provide coverage as great as generics

Taxonomy

- ◆ Tree where each web document is associated with one leaf node
- ◆ Child relationship represents “type of”
- ◆ Available from Yahoo!, DMOZ, etc
- ◆ Includes example documents for leaf nodes

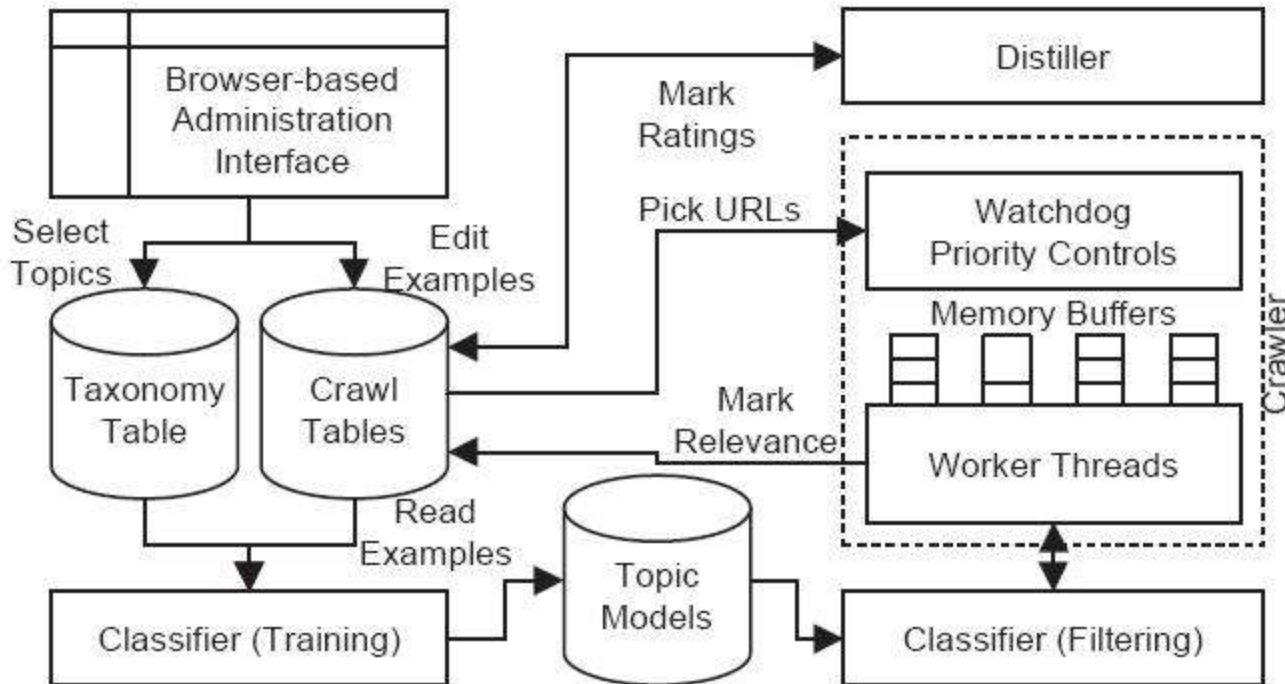
Training A Classifier

- ◆ User provides example documents to be found
- ◆ System recommends potential taxonomy nodes
- ◆ User selects appropriate nodes and verifies relevance of suggested other examples

Training A Classifier (2)

- ◆ Classifier is pre-trained for most classes
- ◆ Classifier re-trains itself for selected classes from example documents
- ◆ Bag-of-words model of document
- ◆ If classes are inappropriate, user may redesign taxonomy

System Architecture



Why Multi-Class?

- ◆ Binary classifier (relevant/irrelevant) is easier
- ◆ Negative examples have little in common with binary classifier
- ◆ Structure of taxonomy may suggest other related classes

Operation Of The Classifier

- ◆ Each document has a probability of belonging to each leaf node.
- ◆ Probability for an internal node is the sum of its children's probabilities
- ◆ Select most probable class
- ◆ True multi-class document is future work

Operation Of The Crawler

- ◆ Begin with citations of example set
- ◆ Hard focus: only follow links from documents whose class or ancestors are relevant
- ◆ Soft focus: follow links from documents whose class or ancestors are relevant first

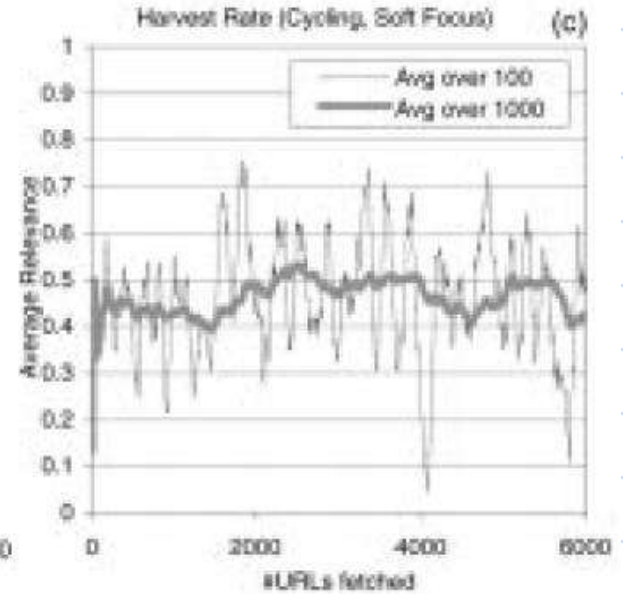
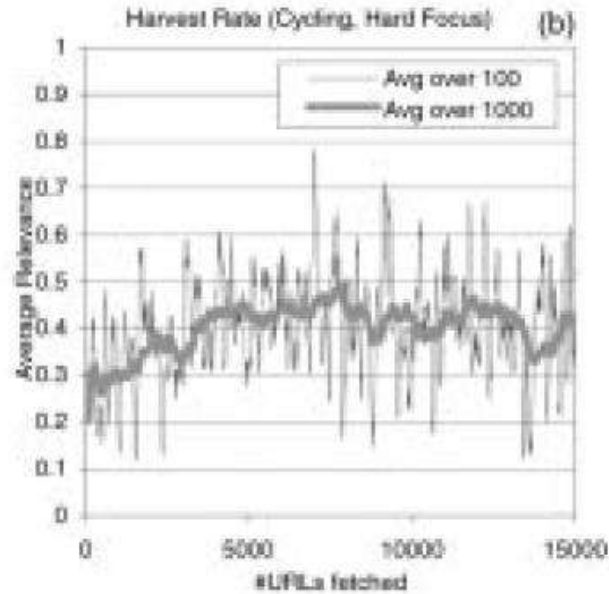
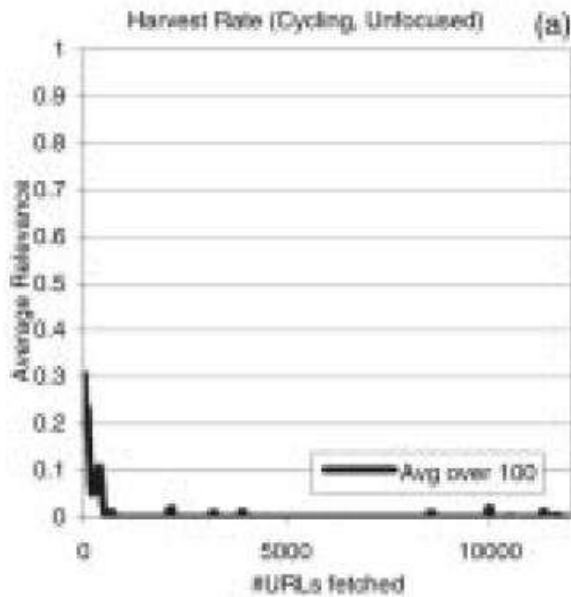
Operation Of The Distiller

- ◆ Prioritize crawling relevant links
- ◆ Based on Kleinberg's hubs and authorities
- ◆ Value of a link to hub and authority scores is related to relevance
- ◆ Only consider authorities with weights above a threshold while iterating

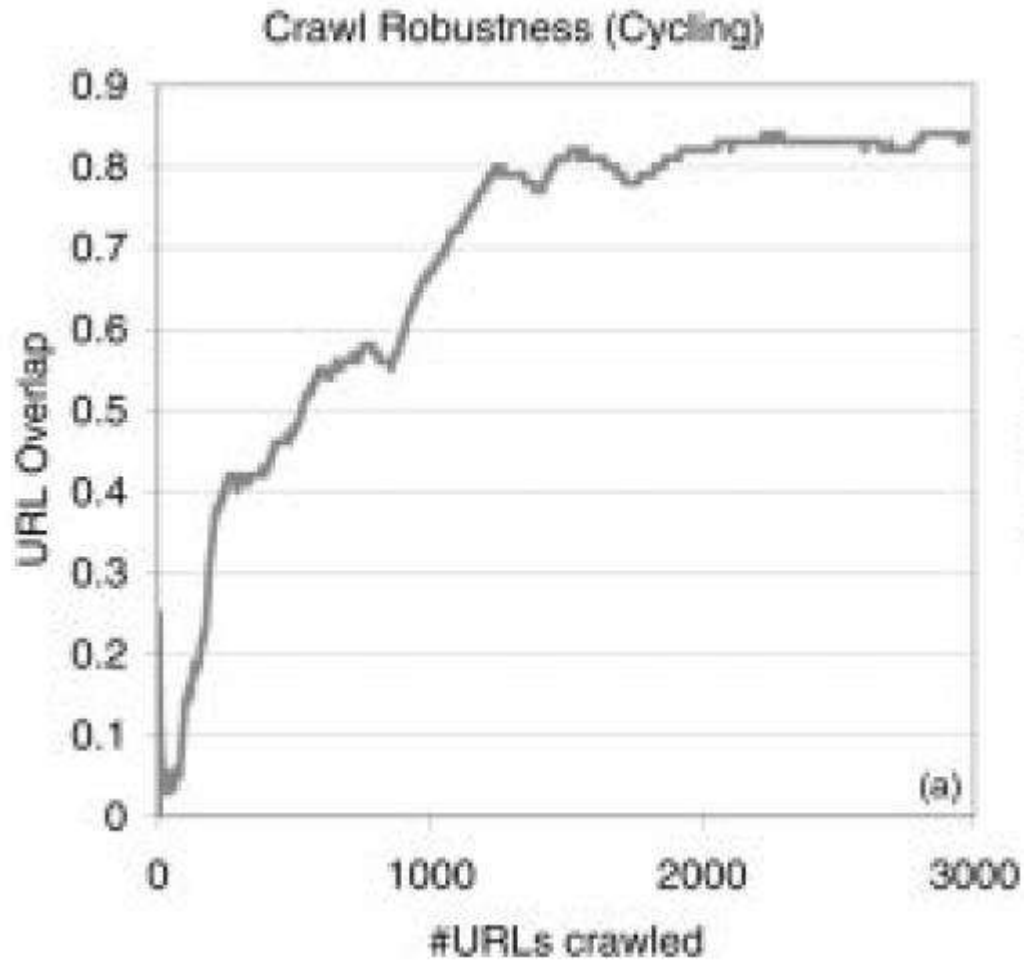
Evaluation

- ◆ Precision, as usual definition
- ◆ Recall is difficult to measure, but starting from different examples finds many of the same pages
- ◆ Harvest Ratio – relevant pages crawled / irrelevant pages crawled

Harvest Ratio



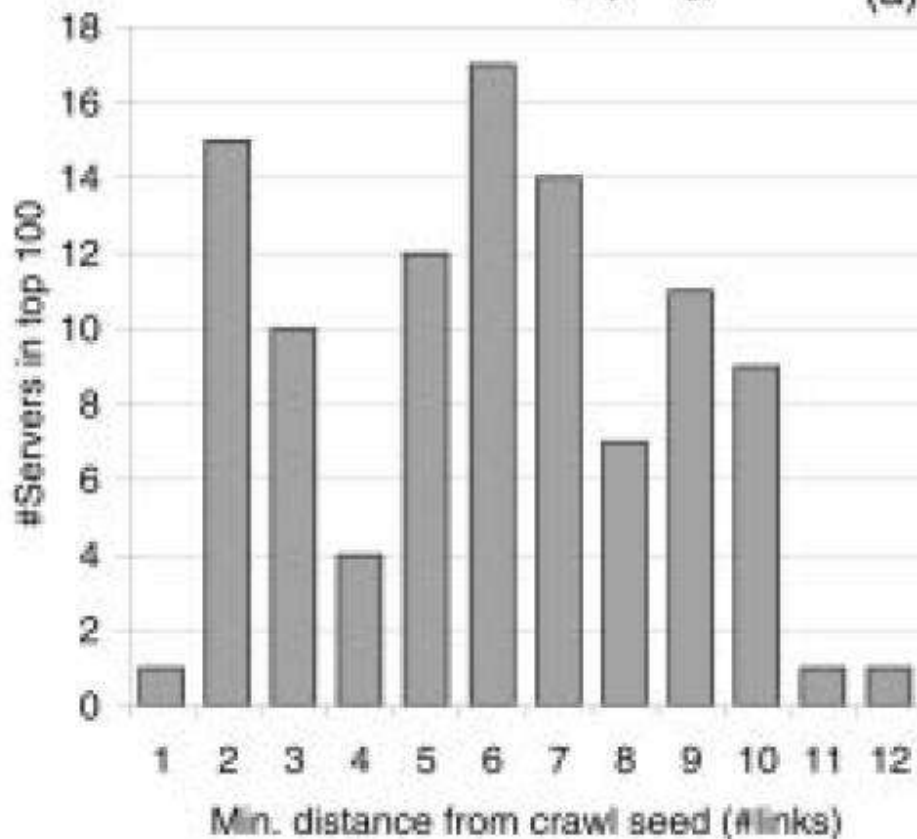
Overlap Of Two Crawls



Citation Distance (100 best authorities)

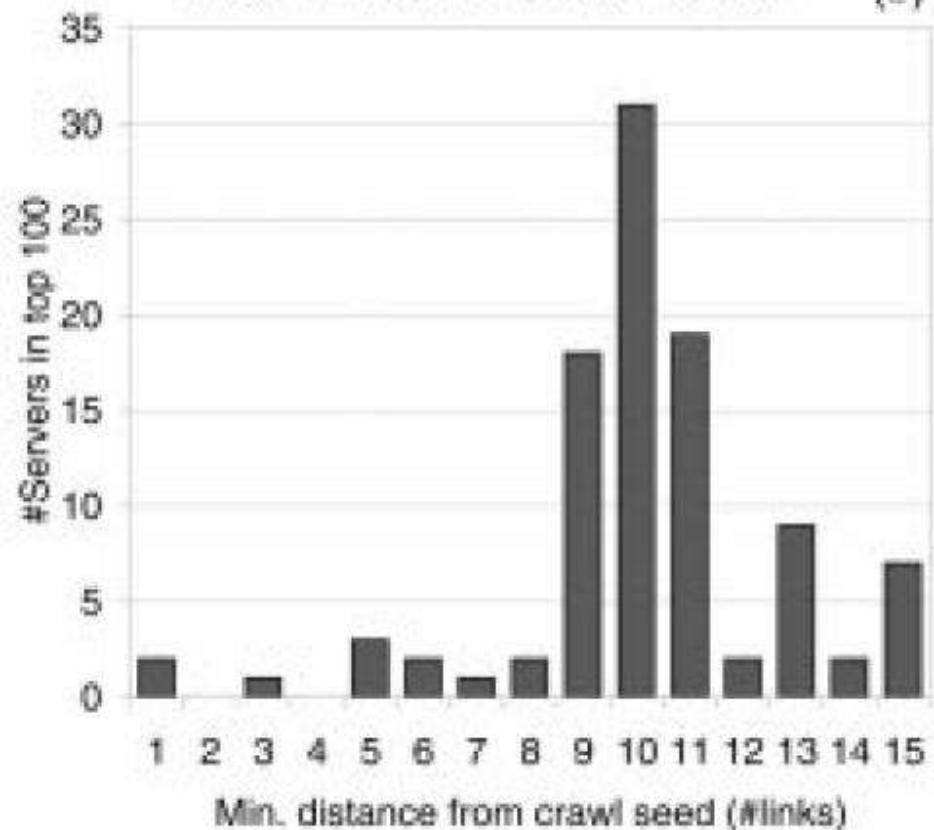
Resource Distance (Cycling)

(a)



Resource Distance (Mutual Funds)

(b)



Questions

- ◆ How large must taxonomy be?
- ◆ Why not run longer, see how hard and soft focus crawling work when most good pages have been found?
- ◆ How is new data incorporated into classifier?