Effects of Start URLs in Focused Web Crawling

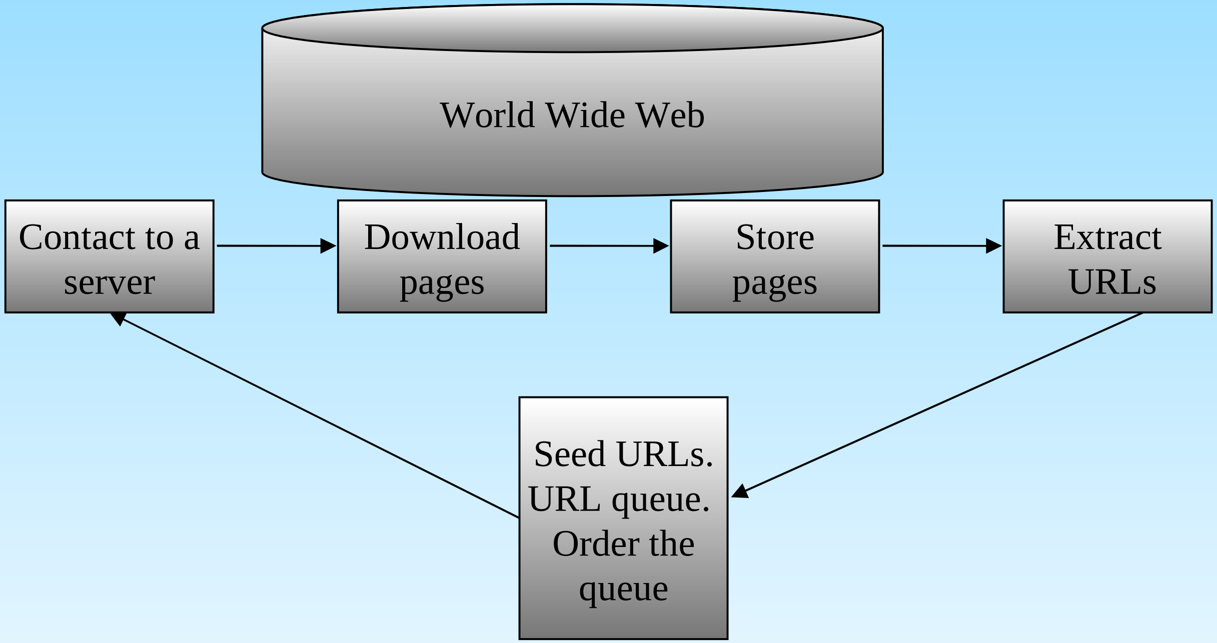
Ari Pirkola and Tuomas Talvensaari

University of Tampere
Finland

Focused crawlers

- Web crawlers: programs that fetch documents (pages) from the Web
- Focused Crawlers *selectively* download Web pages in a specific domain or topic, e.g. genetics, rare diseases, genetic engineering
- Downloaded documents:
    - Domain specific search engines
    - Digital libraries
    - Subject directories
    - Source for data mining

Basic processes of a focused crawler



Research problems

- First Problem: Coverage obtained in different crawling processes started from different geographical regions of the Web
    - Central region (called Major region) contrasted to three other regions: Australia (.au), China (.ch), and five South-American countries (.ar, .br, .cl, .mx, .uy) – these are called Minor regions
        - Major region: (.com, .edu, .gov, .org) and North-American and European countries
    - Coverage = Number of (relevant) documents obtained in crawling

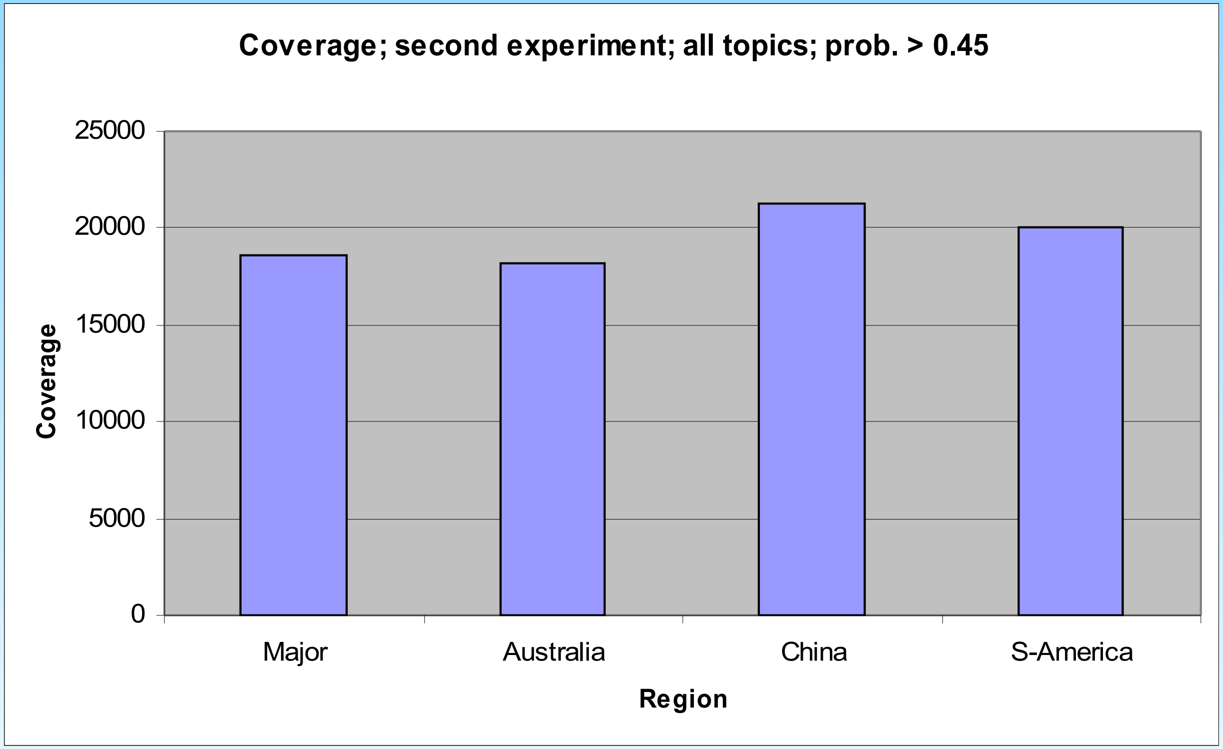
Research problems

- Second Problem: Overlap between the Major region and each Minor region
  - Overlap = Percentage of identical URLs

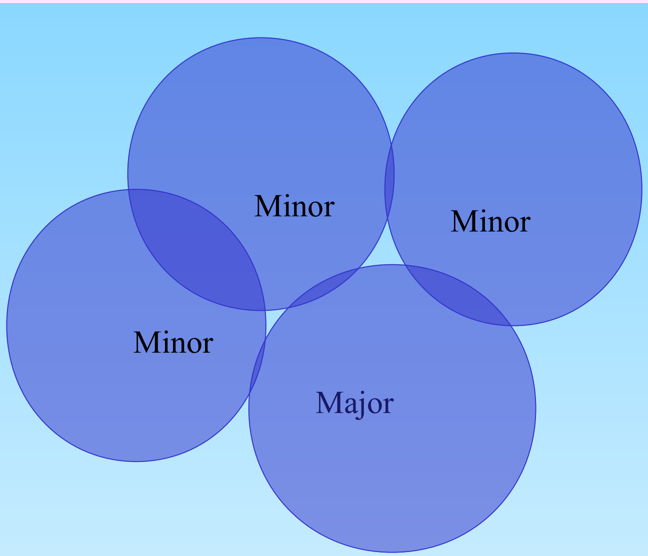
Methods and test data

- 10 test topics in the domains of genomics and genetics
- 50 seed URLs in each case
- Two experiments:
  - First experiment: A text classifier was trained, Terrier search engine, 20 000 pages downloaded for each region
  - Second experiment: Query-document matching, Lemur search engine, 40 000 pages downloaded for each region

Findings


Coverage; second experiment; all topics; prob. > 0.45
25000
20000
15000
10000
5000
0
Coverage
Major
Australia
China
S-America
Region

Findings



circle=relevant documents
obtained for a start region

Overlap rates were low: 0.0%-9.7%

Crawling processes started from different geographical regions identify mainly different relevant documents

Conclusions

- All regions yielded a high coverage
  - - > All are good starting points for focused crawling

- Overlap rates were low
  - - >To be able to collect a large topic-specific document collection one has to use different start URL set

Conclusions – future research

One key issue for future research is to investigate how to obtain large topic-specific document collections (e.g. for digital libraries). In addition to different starting points, one could use different crawling methods. For example, we have devised a focused crawler that identifies equivalent link words in different languages on the basis of fuzzy matching (e.g. English *genetic* and German *genetisch*), as well as variant forms of the same basic word within a given language (e.g. *mutation, mutant, mutate*). It seems that the only way to obtain a high coverage in topic-oriented focused crawling is to combine the results of different approaches (starting points, methods).