The Web IR - Challenges

Vasudeva Varma

IIIT Hyderabad

Focused Crawling

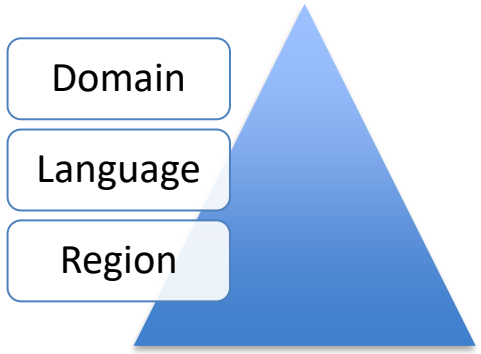
Agenda

- Vertical Search
- Focused vs unfocused crawling
- FC Strategies/Applications/Benefits
- FC Metrics
- Types of FCs
- Seed selection in FCs
- Diversity in Seed set

Domain Specific Search

Success Factors

- Understanding Country's languages and culture
- Region and domain specific tailoring of the technology
- Three dimensions of "useful" search technology




Baidu 百度
Search
Web
General
Image
Regional Search
Movie
Map
Ancient Literature
Government
Youth Search
Education
Video Search
Book Search
MP3
News
Mobile Search
File Search
Mobile Entertainment
Yellow Pages
Postal Codes
Legal
University
Blog Search
Web Search
Culture Search
Community
Post Bar
WAP Post Bar
WAP Knows
RSS News Feed
Baidu Knows
Baidu Greetings
Baidu Encyclopedia
Baidu Space
Music Box
Mobile Entertainment
Olympic Channel
Game Channel
Enhancements
Sobar
Dictionary
Top of the Charts
Desktop Search
Search Index
Music Master
Baidu Favorite
Anti-Virus Platform
Download Bar
Industry Report

Domain Specific Search Is the Next Big Market

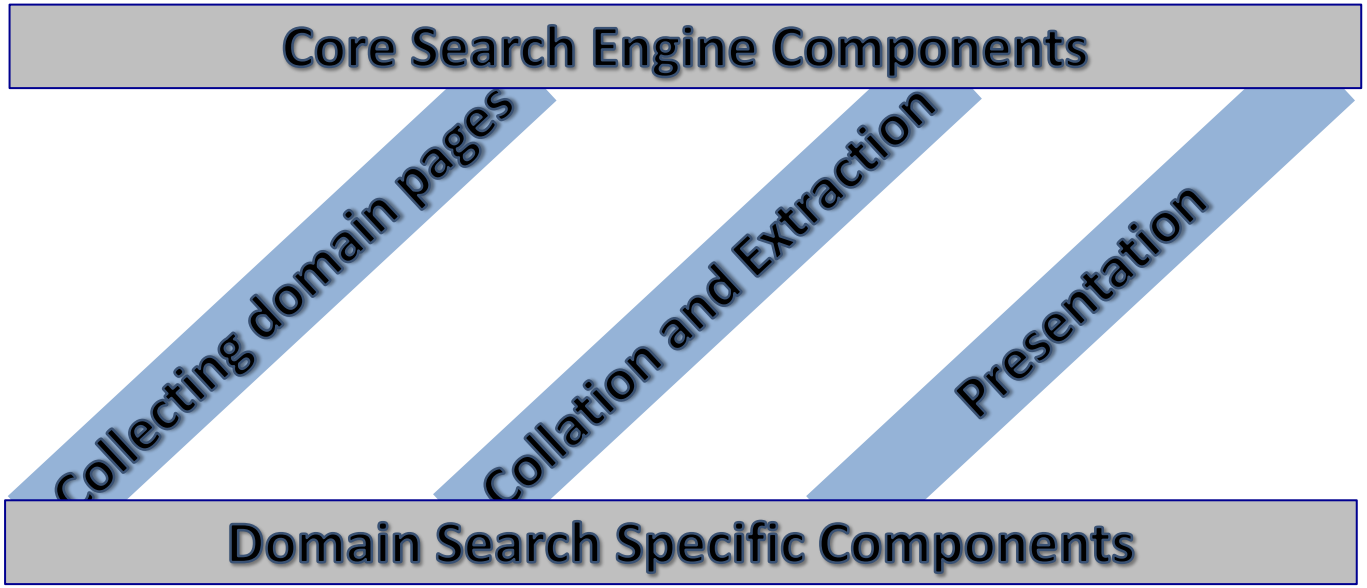
- Vertical Search and Semantic Technology Will Change Digital Advertising

- Pay-Per-Click campaigns with a vertical search engine result in higher clickthrough rates and higher conversion rates

Curious case of Qunar:

A Proven Model of A Vertical/Domain Search Plus A Transaction-enabled Service

Solution: Components of the platform

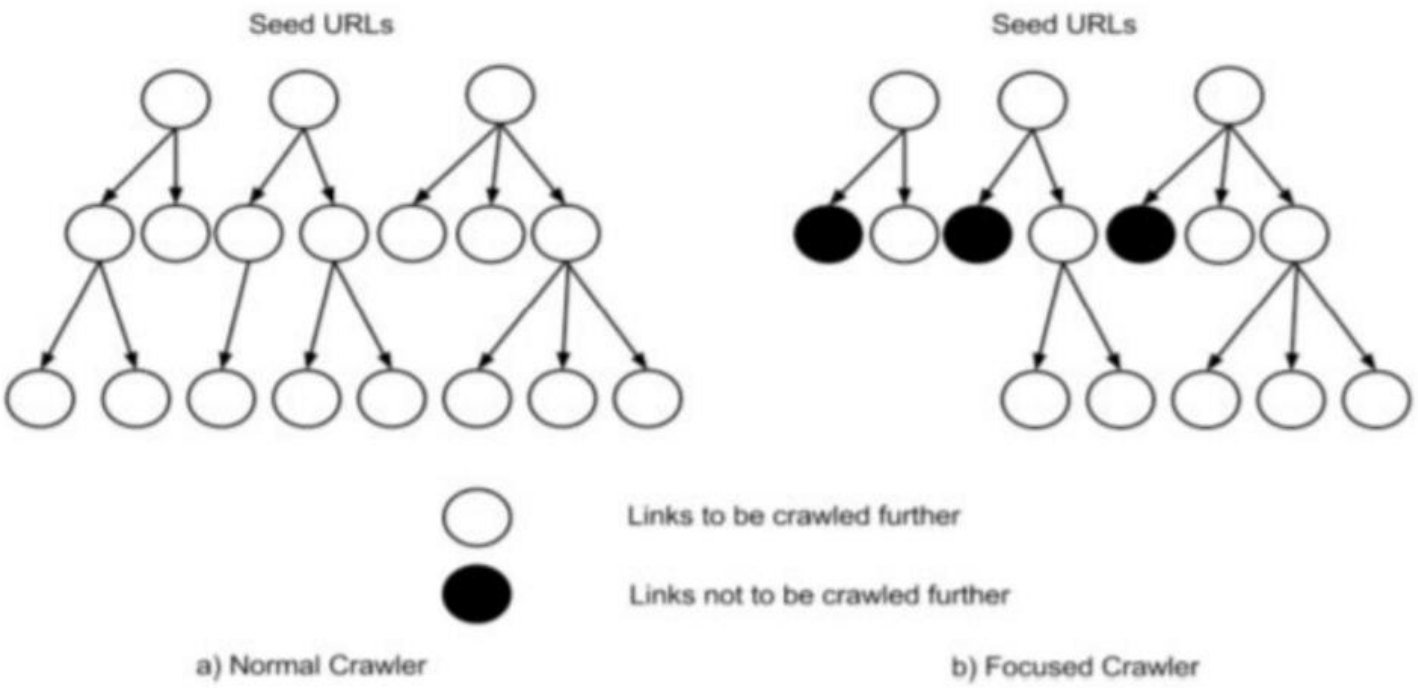


System Components – Core and Domain

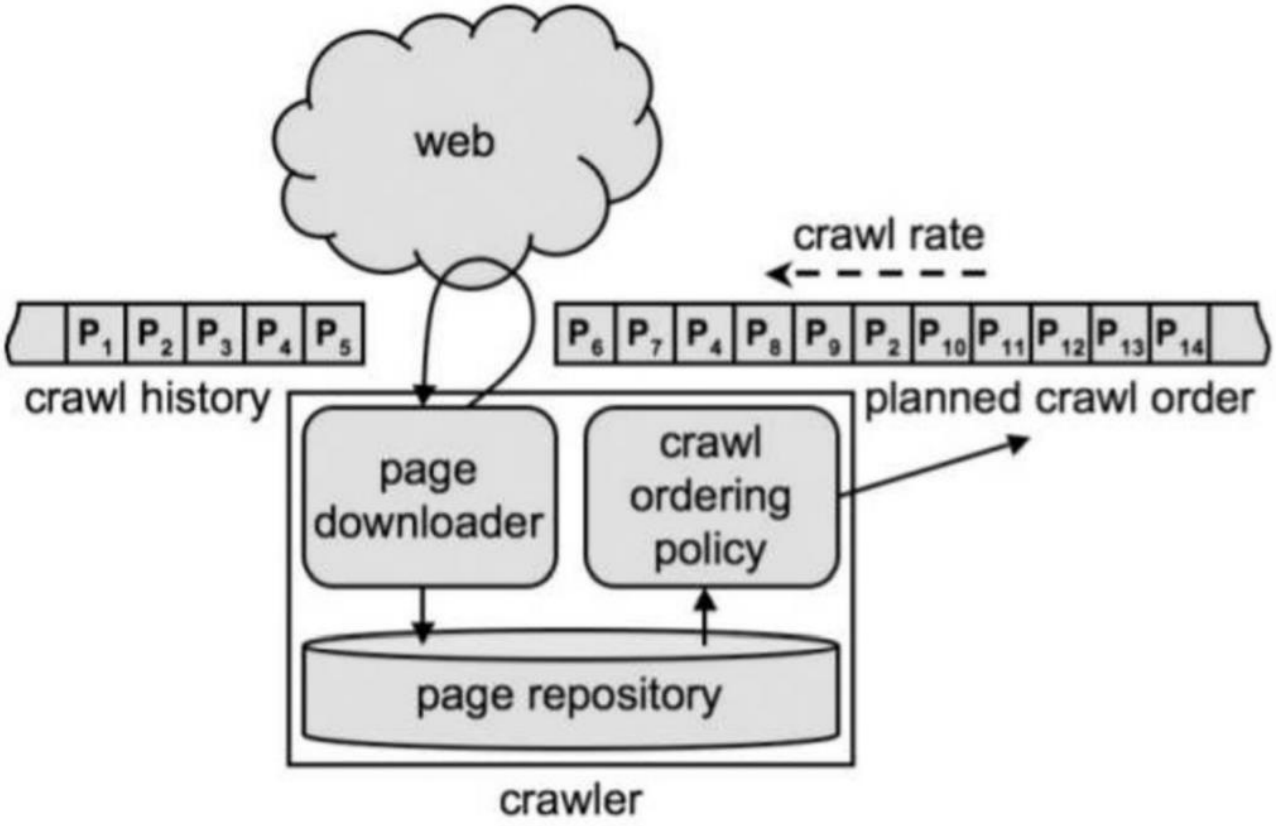
- Crawler
- Document Parser
- Query Parser
- Indexer
- Search Module
- Ranking Module
- User Interface (Web and Mobile)
- Evaluation

- Building seed sets for each domain
  - Domain Crawling and Classification Models
  - Domain Dependent Parsers

Focused vs unfocused crawling: Identify the links not to be crawled vs crawling all links



Simplified crawler architecture



Crawling strategies

- Crawl pages from only the .in domain
- Crawl pages with large PageRank
- Crawl pages about Cricket
- Crawl pages written in ‘Hindi’ language

Applications

- Building domain or genre specific search engines
- Building Personalized Search Tools
- Extending digital libraries
- Discovering linkage sociology
- Locating specialty sites (E.g. specialized sites for mountain biking)
- Acquiring training data for ML tasks
- Detecting community culture
- Estimating community timescales

FC Metrics

- Precision: No. of relevant pages in crawl / Total number of pages crawled
- Recall: No. of relevant pages in crawl / Total number of relevant pages in the entire web
- Harvest ratio: Rate of change of precision per unit time.

Benefits of FCs

+ Results are more relevant

+ Saves resources such as time/space/computational power/bandwidth

+ Fresh crawl

- Trade off is recall

Types of FCs

- Early Algorithms: Fish Search & Shark Search
- Classifier based focused crawlers
- Reinforcement learning based crawlers
- Context graph based focused crawlers
- Ontology based focused crawlers
- Page properties based focused crawlers
- Incrementally learning crawlers
- Adaptive Crawlers
- Ant based crawlers

Seed selection for Focus Crawling

- Blacklist approach: Names of hosts to be avoided
  - Blacklist has to be created manually
  - As the blacklist grows longer, so does the time the crawler spends on blacklist filtering.
  - Start with any set of seed URLs which are not blacklisted
- Whitelist approach: start the focus crawl from a list of high-quality seed URLs and limit the crawling scope to the domains of these URLs
  - Is generated using long crawling history
  - Must be updated periodically
  - URLs in the whitelist should provide the majority of useful documents.
  - URLs should be sorted based on their rank indicators so that top ranked URLs have the highest priority to be crawled.

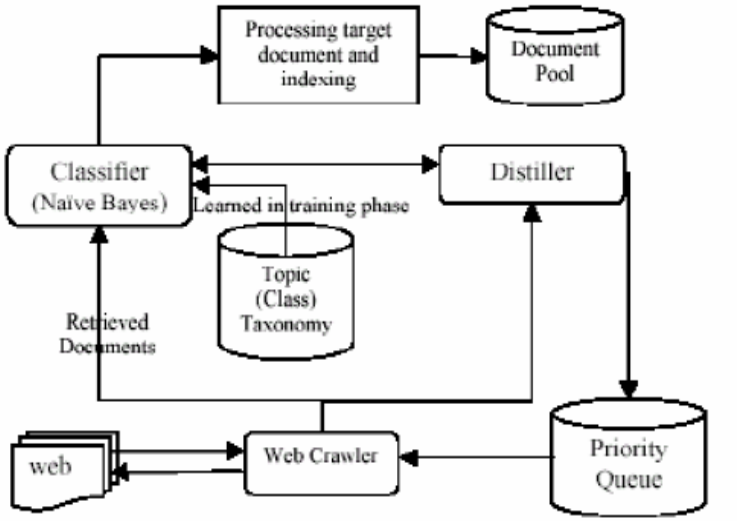
Diversity in Seed set

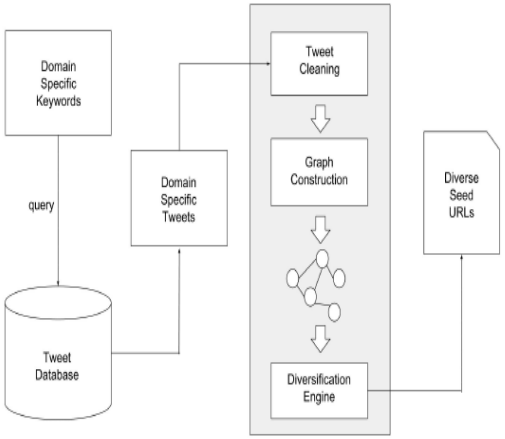
Even though we wish our seed set to be focused on a particular domain we want it to be diverse within that domain (Why?)

Consider the domain "tourism", we would want to have the seed set containing pages of all sub-domains:

- Hotels
- Transportation
- Places of historic significance
- Famous Cuisines
- Shopping Malls
- Weather reports
- Tourism Industry
- Tourism Ministry
- Travel or VISA related information

Domain Specific Crawler





Selecting seeds for domain

Key messages

- It takes time to focus
- It takes effort and time to remove junk than to get the right pages
- Achieving high recall is the main challenge

Further reading

- Introductory article on deep web link
- Section on crawling the deep web, from this university's guide on deep web:
- Optional Reading: Paper on crawling the deep web
- Victor Lavrenko short videos on web crawling
- Implementation view