ProjFocusedCrawler
CS5604 Information Storage and Retrieval, Fall 2012
Virginia Tech
December 4, 2012

Mohamed M. G. Farag
Mohammed Saquib Khan
Prasad Krishnamurthi Ganesh
Gaurav Mishra

Outline

- Project description
- Deliverables
- Architecture
- Roles
- Progress
- Milestones
- Problems and challenges

Project Description
(from course Scholar webpage)

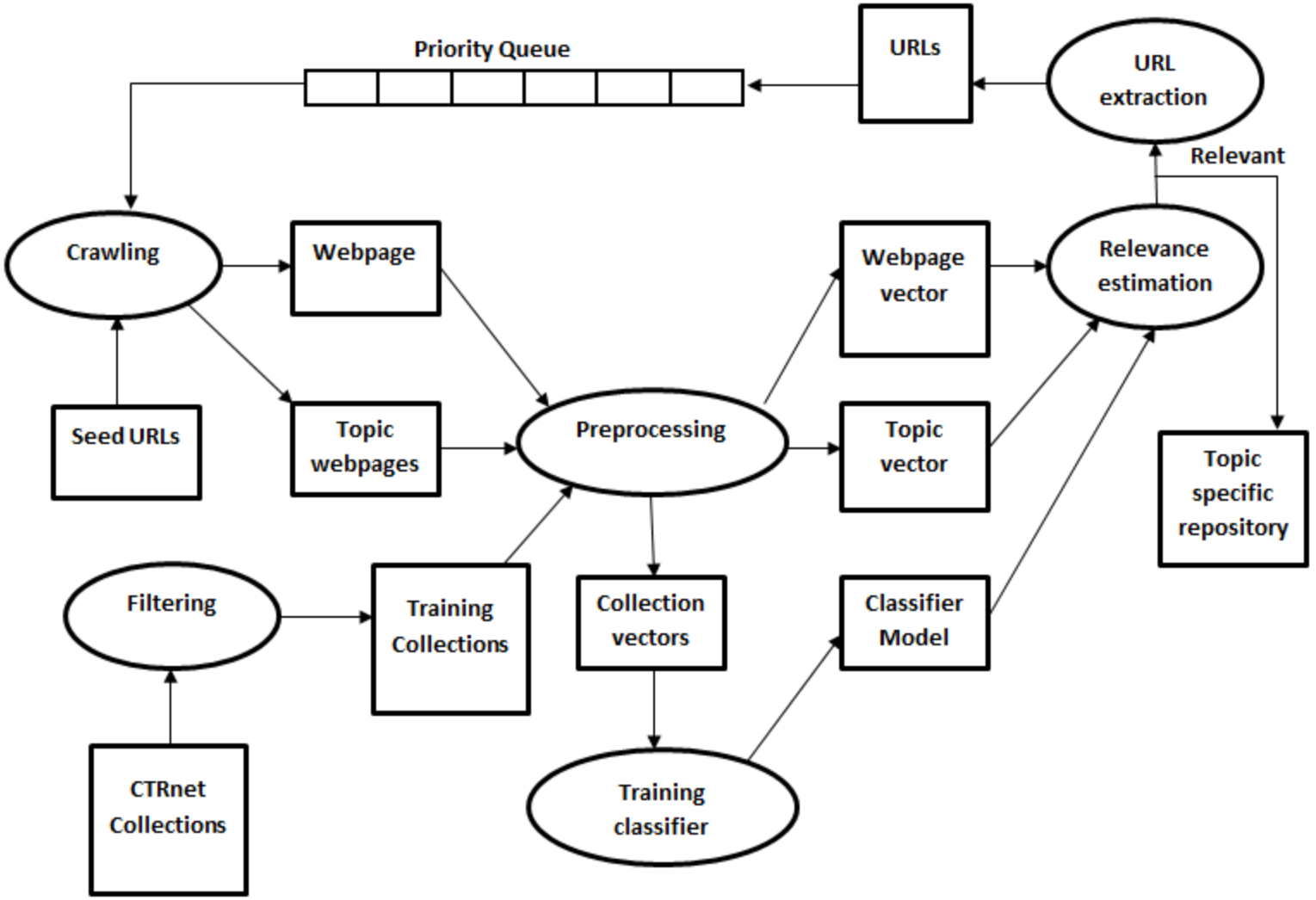
- CTRnet uses Heritrix
  - Quality of the seeds, and configuration details
- Focused crawlers
  - use topic information (which links to follow)
  - reduce noise, and increase precision
  - may reduce recall
- Goal
  - improve upon existing solutions
  - demonstrate effectiveness w.r.t. CTRnet efforts

Deliverables

(from course Scholar webpage)

- Design and implementation of a prototype focused crawler
- New collection for CTRnet built using crawler
- Software, data, report, and future plans
- Publication about this research, and content for NSF proposal

Architecture



Roles

- Mohamed M. G. Farag:
  - Building classifier
  - Preparing training data (lead)
  - Integrating different components
- Mohamed Saquib Khan:
  - Building classifier (lead)
  - Preparing training data
  - Comparing different classifiers

Roles (cont'd)

- Prasad Krishnamurthi Ganesh:
  - Preprocessing (lead)
  - Preparing training data
  - Evaluation
- Guarav:
  - Preprocessing
  - Preparing training data
  - Building documents TF-IDF vectors (lead)

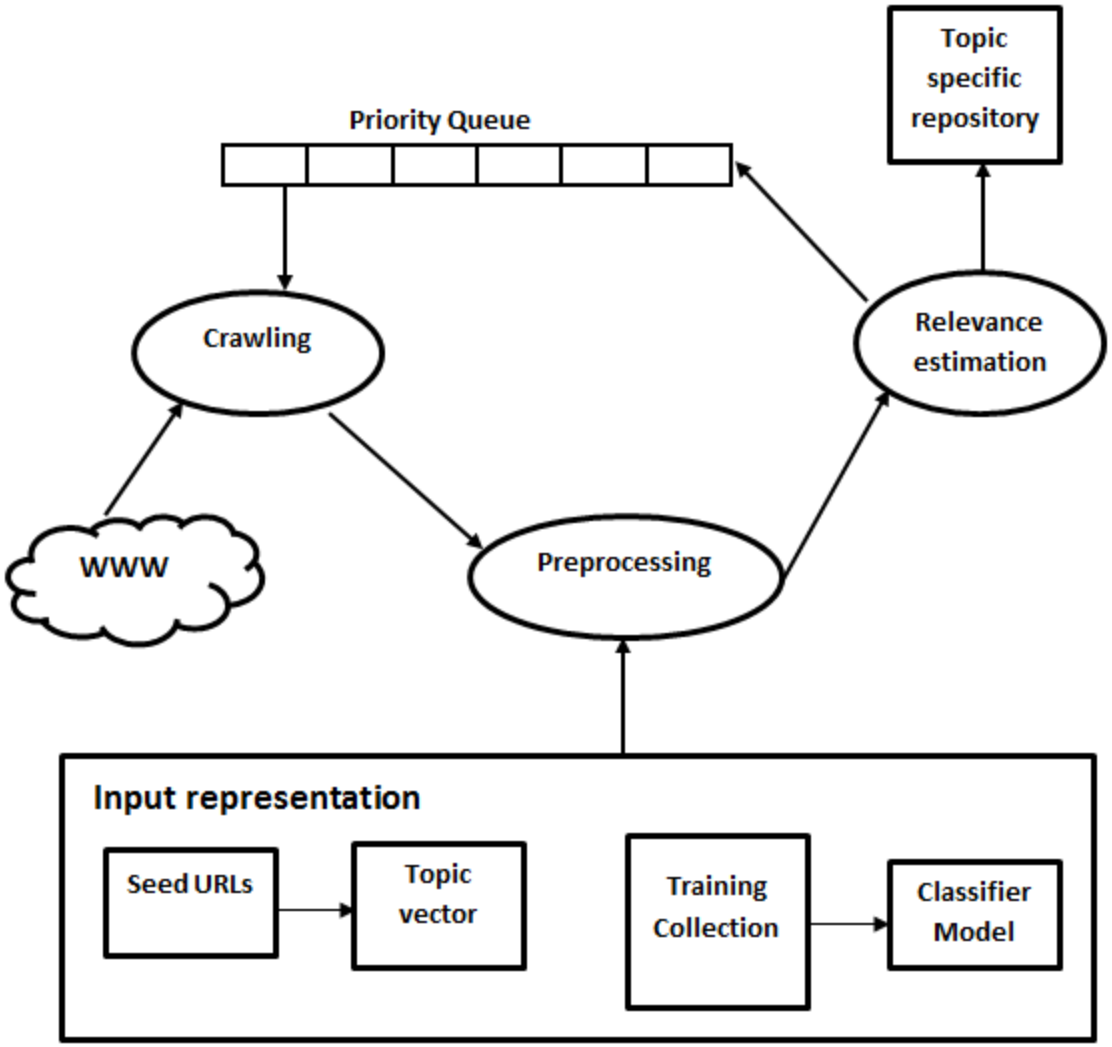
Preparing training data

- Warc files extraction
- Sikkim earthquake collection
  - 19 seed URLs
  - ~2000 HTML files out of 9000 files
- Filtering
  - Keywords, selected manually
  - Relevance -> k or more words from keywords
  - K = 1 → ~50% relevance (high recall, low precision)
  - We used k = 5

Problems faced

- Warc extraction
  - Files saved on disk, original URL, wayback machine URL
  - Encoding problems of saved on disk files
  - Original URLs, not working (page not found)
  - Wayback machine URLs, webpages have extra content that needs to be parsed and removed

Modular design



Input representation

- Seed pages
  - Vector space model
  - Normalized TF weighting scheme
    - Problem of using IDF of relevant documents only
- Training and testing data set
  - Features, vector space model
  - TF-IDF weighting scheme

Crawling

- Get a URL from priority queue
- Check if URL is:
  - Visited (i.e., its webpage is already downloaded)
  - In Queue (i.e., no need to put it again in queue)
- Download
- Extract text and URLs
- Estimate relevance
- Put URLs in priority queue
- Stop if queue is empty or reach pages limit

Text extraction problems

- Extract visual text of a webpage
- Heuristic approaches
- Scripts and comments tags remain after extraction (need explicit manipulation)
- Invalid HTML tags (missing closing brackets)

How to Crawl

- Three ways:
  - URLs only: anchor text and address (doesn't describe the topic of target webpage)
  - Webpage text only: time and space, all URLs get same score, many non relevant URLs
  - Hybrid: if webpage is relevant, use relevant URLs only

Relevance estimation

- TF-IDF/cosine similarity score
  - Assumed relevant if score more than threshold
- Classifier
  - Naïve Bayes
  - Support Vector Machine (SVM)

Evaluation

- Precision
  - Number of downloaded webpages that are relevant
- Classifier evaluation
  - Performance on test data
  - Cross validation, parameters best values
- Performance
  - Ordering of URLs in priority queue

Results

- Egyptian revolution, threshold = 0.1
- Precision = 633/1208 = 0.52

URL	Score
http://botw.org/top/Regional/Africa/Egypt/Society_and_Culture/Politics/Protests_2011/	1
http://live.reuters.com/Event/Unrest_in_Egypt?Page=0	1
http://www.aljazeera.com/indepth/spotlight/anger-in-egypt/	1
http://www.guardian.co.uk/world/series/egypt-protests	1
http://www.huffingtonpost.com/2012/06/24/egypt-uprising-election-timeline_n_1622773.html	1
http://www.washingtonpost.com/wp-srv/world/special/egypt-transition-timeline/index.html	1
http://www.guardian.co.uk/news/blog/2011/feb/09/egypt-protests-live-updates-9-february	0.52428340
http://www.guardian.co.uk/world/blog/2011/feb/05/egypt-protests	0.50552904
http://www.guardian.co.uk/world/blog/2011/feb/11/egypt-protests-mubarak	0.50212776
http://www.guardian.co.uk/news/blog/2011/feb/08/egypt-protests-live-updates	0.47775149

SVM

- Without feature selection, k = 5 for filtering
- Accuracy = 0.845132743363

	precision	recall	f1-score	support
0	0.85	1.00	0.92	191
1	0.00	0.00	0.00	35
avg / total	0.71	0.85	0.77	226

SVM

- With feature selection (Chi-square) and k = 5 for filtering
- Accuracy = 0.898230088496

	precision	recall	f1-score	support
0	0.89	1.00	0.94	191
1	1.00	0.34	0.51	35
avg / total	0.91	0.90	0.88	226

Naïve Bayes

- Same result for using/not using feature selection and k = 5 for filtering
- Accuracy = 0.845132743363

	precision	recall	f1-score	support
0	0.85	1.00	0.92	191
1	0.00	0.00	0.00	35
avg / total	0.71	0.85	0.77	226

Future work

- Domain ontology
- Tunneling
  - Non-relevant webpage can lead to relevant ones
- Ontology-based relevance estimation
  - Concepts comparison

Milestones

Milestone	Date
Preparing training data	(Oct. 15)
Building general crawler	(Oct. 30)
Building classifiers	(Nov. 15)
Integrating components	(Nov. 30)
Testing and evaluation	(Dec. 5)
Completing report and prototype	(Dec. 11)

Thank you!

Questions ?