Mining the Web for Multimedia Content

Brett R. Gordon
Carnegie Mellon University
Language Technologies Institute

August 19, 2002

Overview

- Problem definition
- Requirements
- Assumptions
- Why this problem is difficult
- Methods/Results
- Suggestions for future work

General Problem Definition

- Large amount of data exists, in non-textual form, distributed over the Web
  - Tables, charts, diagrams, etc.
- Usefulness of this information varies
  - Isolated pieces may have no value
- Aggregating this information can be useful
- Goal: Build Question Answering (QA) system with multimedia content
  - Increase scope of possible questions
  - Increase richness of possible answers

Specific Problem Definition

- Find distributed web pages about a certain domain (i.e. agriculture) that contain a specific type of media (i.e. tabular)

Requirements

- Define a domain
  - Distributed data must exist (non-centralized)
  - Data should exist in sufficient quantity
  - Address temporal nature of the data
- Select a media type
  - Current focus is on tabular data
- Establish a relevance decision
  - Given a page → { relevant | non-relevant }

Assumptions

Implicitly assumes that data in target domain of a specific type

1. Exists (in a readable format)
2. Not isolated
3. Not centralized in an unknown source
   - If data exists in a centralized form, then there is no value-added to the system

Why this problem is difficult

- Difficult search space
  - Web is very large
  - Highly noisy
  - |{Relevant Docs}| < |{Non-Relevant Docs}|
- Difficult relevance decision
  - Little or no labeled data
  - Many web pages contain non-relevant tables
    - But this can be difficult to determine

General Approach

- Start from a set of seed pages
- Breadth-first search (a.k.a. spidering)
  - Assumption: Target pages can be reached from set of seed pages
  - Assumption: Target pages are in close proximity to seed pages

Seed Pages

- Pros
  - May significantly reduce search time
  - May improve precision/recall
  - Manual effort required to identify seed pages may be minimal

Seed Pages

- Pros
  - May significantly reduce search time
  - May improve precision/recall
  - Manual effort required to identify seed pages may be minimal

- Cons
  - May ***not*** significantly reduce search time
  - May ***not*** improve precision/recall
  - Manual effort required to identify seed pages may be ***substantial***

Approach #1: Bare Bones

- Manually selected 20 seed pages
- Manually selected 30 keywords
- Relevance decision
  - Page contains **at least** one keyword
- Results
  - Total Pages Crawled: 18,000
  - # Possibly Relevant: 148
  - # Actually Relevant: 7

Approach #2: Added Sub-Tree Threshold

- Updated keywords and seed pages
- Relevance decision (same)
  - Page contains at least one keyword
- Threshold decision
  - Do not follow any more links along a path if X links have been followed without any relevant documents
- Results
  - Total Pages Crawled: 50,000
  - # Possibly Relevant: 447
  - # Actually Relevant: 13

Approach #3: Added table ratios

- Relevance decision
  - Page contains at least one keyword

    AND
  - Page contains a table with at least 50% numerical tokens
- Results
  - Total Pages Crawled: 1.3 million
  - # Possibly Relevant: 1,672
  - # Actually Relevant: 108

Comments

- No significant change in results across methods
- Exhausted the relevant pages in this search space
  - Within a certain neighborhood of the seed pages
- Must reduce the search space

New Approach

- Focus so far has been at the page level
- More useful to focus at the domain level?
  - Some domains are more/less relevant than others
- How can we build a stop-domain list?
- How do we decide whether a domain should be added to the list?

General Crawl

- Step 1
  - Given a media type and some seed pages, what topics can be found?
    - Use seed pages
    - No keyword list
    - More exploratory objective
  - Crawl

General Crawl

- Step 2
  - Estimate the **expected** # of relevant pages in a domain
    - Use Google to estimate size of a domain
    - Use current sample estimate
  - Remove domains
    - That have already been mostly crawled
    - That have had a minimum number of pages visited and a large number remaining
  - Keep domains
    - That have a **high expected relevant page return**
    - That have not been sufficiently crawled to decide

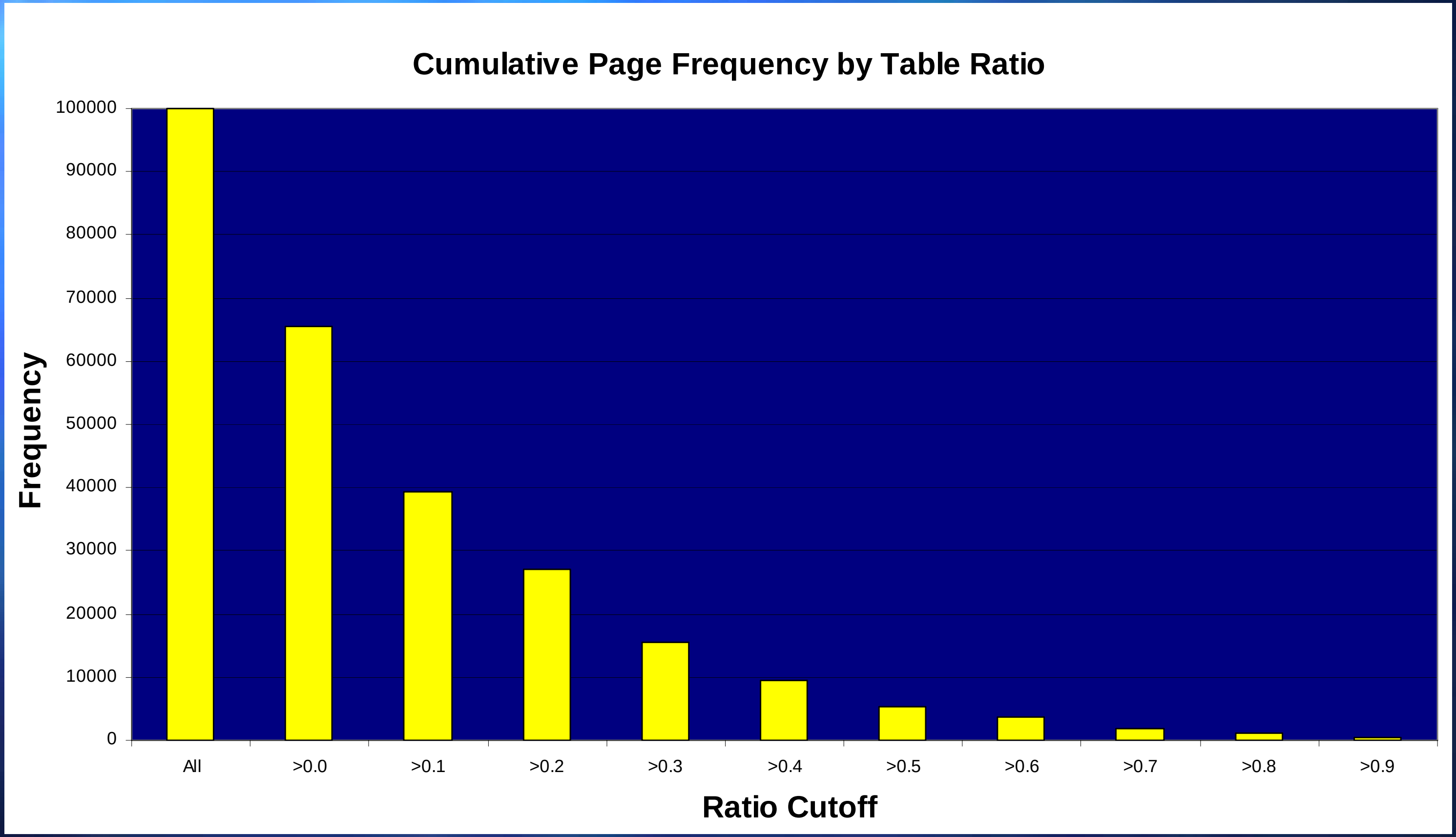
Domain Filtering Results

- Total pages crawled: 100,000
- Unique domains: 7,300
- Filtering removes 955 domains
  - 1 domain was actually useful
  - 954 were not
- Reduced remaining search queue from 917,000 to 175,000

Table Ratio: How Useful?

- Relevant tables will have a high ratio
- Many non-relevant tables may also have a high ratio
  - Especially in small tables
- What is the distribution of table ratios?

Distribution of Table Ratios



Page Evaluation

- Divided corpus into 10 bins according to table ratio
- Randomly sampled from each bin
- Manually classified samples

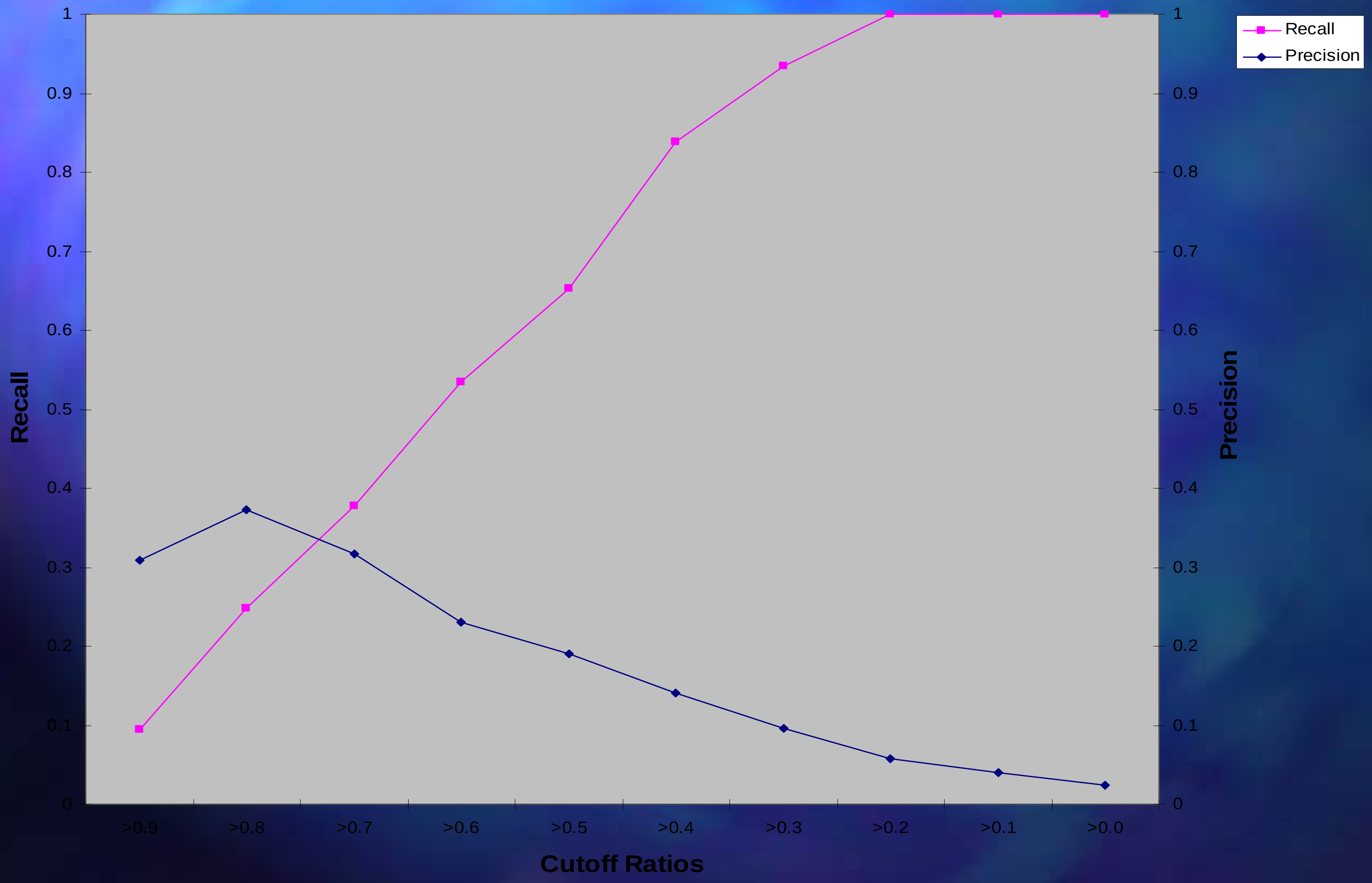
Page Evaluation: Results

Bin	*Frequency*	*Sample Size*	*# Relevant (sample)*	**Expected # Relevant (pop)**	*% of Bin*
[0.0 - 0.1)	27068	200	0	0	0
[0.1 - 0.2)	12955	200	0	0	0
[0.2 - 0.3)	10377	200	2	104	0.010
[0.3 - 0.4)	6099	200	5	152	0.025
[0.4 - 0.5)	3638	100	8	291	0.080
[0.5 - 0.6)	1868	100	10	187	0.100
[0.6 - 0.7)	1651	100	15	248	0.150
[0.7 - 0.8)	860	100	24	206	0.240
[0.8 - 0.9)	535	100	45	241	0.450
[0.9 - 1.0]	485	100	31	150	0.310

Page Evaluation: Cumulative Results

Cutoff	# of Pages	Cumulative Expected # Relevant	Precision	Recall	Absolute Return
>0.0	65598	1579	0.024	1.000	0.016
>0.1	39245	1579	0.040	1.000	0.016
>0.2	27188	1579	0.058	1.000	0.016
>0.3	15460	1475	0.095	0.934	0.015
>0.4	9437	1323	0.140	0.838	0.013
>0.5	5399	1032	0.191	0.654	0.010
>0.6	3675	845	0.230	0.535	0.008
>0.7	1880	597	0.318	0.378	0.006
>0.8	1050	391	0.372	0.248	0.004
>0.9	485	150	0.309	0.095	0.001

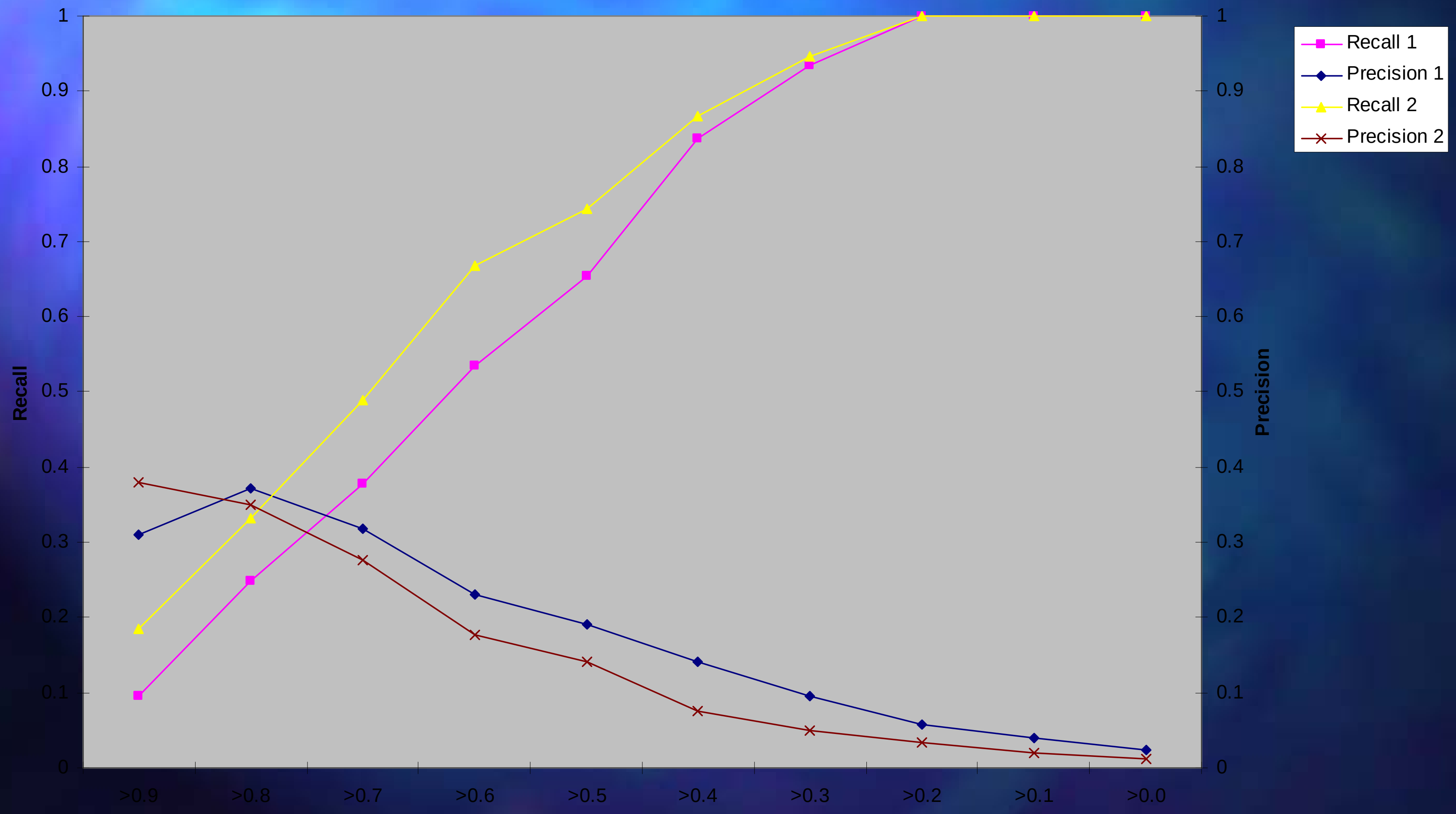
Precision/Recall



Second Crawl

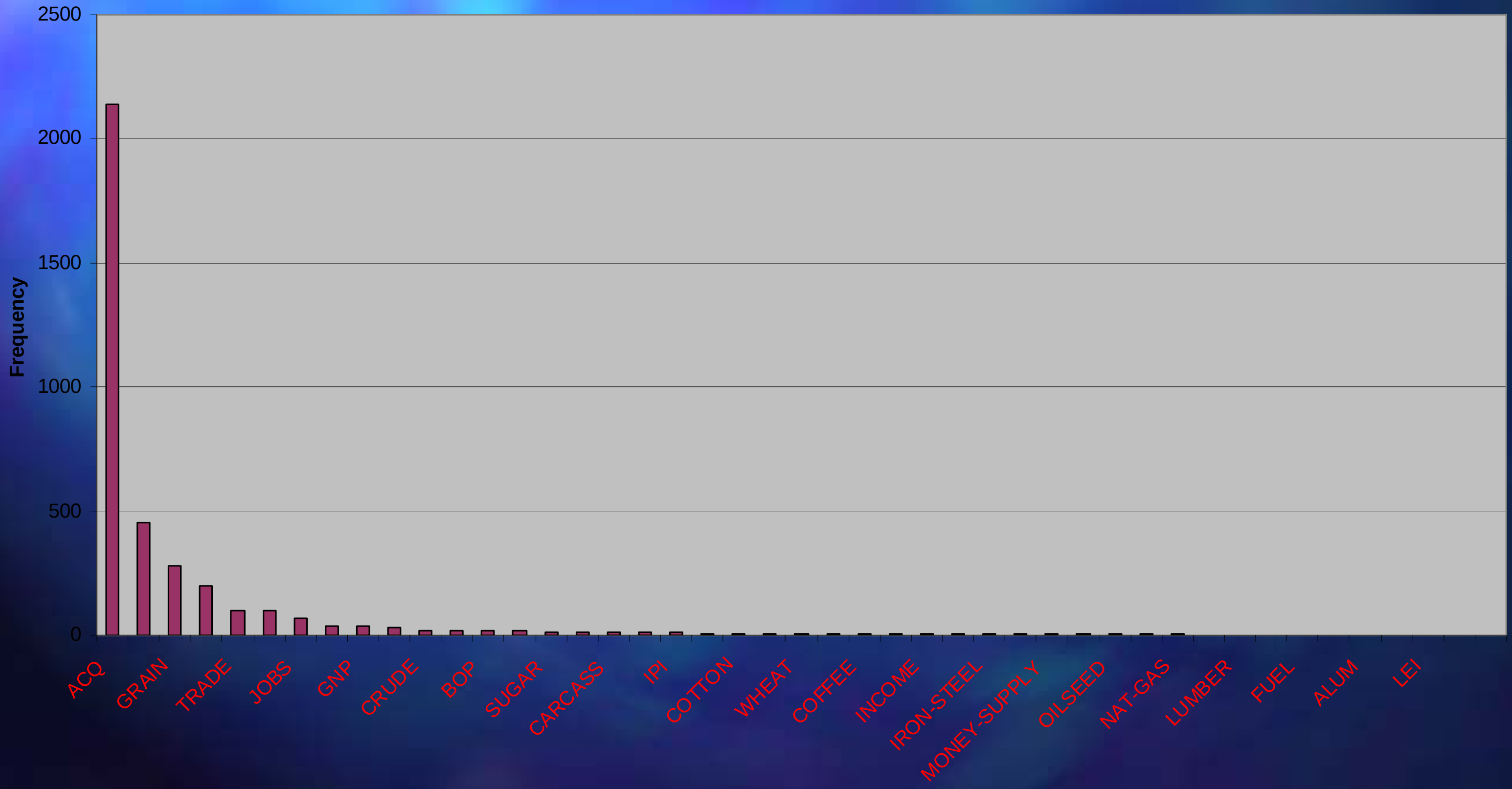
- Re-ran crawler using new stop-domains
- Total Pages Crawled: 100,000
- Re-sampled and manually labeled

Precision/Recall: Second Crawl



Classification

- Classified pages above 0.6 cutoff ratio (3200 pages)
- SVM trained on Reuters 21578



Suggestions for Future Work

- Effective filtering is as important as effective identification
  - Need more efficient and effective means to identify non-relevant domains
  - Better page-level relevance judgments
    - Identify/Parse numerical tables more accurately
- Add user feedback into search process
  - Domain and page level feedback
  - Add keywords from newly crawled pages

Suggestions for Future Work

- Other information sources
  - Find ‘similar’ documents through Google
  - Incorporate external domain information (Google, DMOZ, etc.)
  - Google API
    - http://www.google.com/apis