Course-specific Search Engines: Semi-automated Methods for Identifying High Quality Topic-specific Corpora

Neel Guha, Matt Wytock
Gunn High School, Carnegie Mellon University

May 13, 2013

The educational context

- Web search is an important tool for students of all levels
- In this work, we focus on the high school level and specifically the AP United States History (APUSH) course taken by 400,000 students in 2011
- For these students, there are a number of problems with generic keyword-based search engines
- Goal: Create a search engine specialized to every course

Problems with keyword search

- **Off-topic results**. [benjamin franklin] brings up plumbing service; [gold rush] brings up pages related to Gold Country tourism

- **Inappropriate sources**. Many sources not reputable, including user-generated content (forums, Yahoo answers); sites offering other student essays; biased sites (ConfederateAmericanPride.com)

- **Wrong level**. [thomas jefferson] returns a page from the children's version of Library of Congress; typically no explicit labeling for level

Google Custom Search Engine

- Takes care of difficult task of crawling the web, building an index and running a search engine
- Given a list of sites, can boost these in the results or restrict to just these sites (we use restrict in our experiments)
- Available at `http://google.com/cse`

APUSH textbook

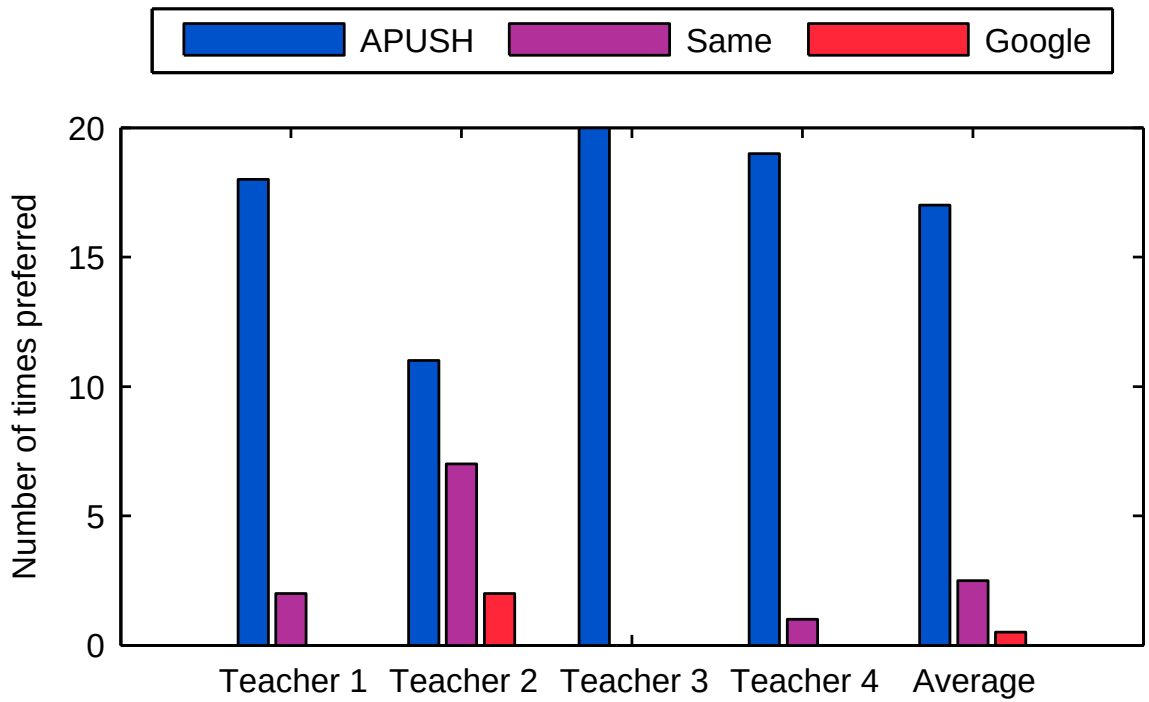
- The American Pageant, Twelfth Edition. (available online)
- 42 chapters, 1034 pages
- Authoritative document describing the course
- Structured information such as chapters, sections, charts, tables, chronology, references, etc. (not yet used)

APUSH reference search engine

- We extracted 1206 distinct proper nouns using syntactic cues (capitalization, puctuation, etc.) and took combinations of these to form our query set
- We issue these queries to Google and extract 23393 sites with 1757 occurring $>$10 times
- Manually curate list of sites (56% good) and use this to build the APUSH CSE

APUSH CSE evaluation

- Blind side-by-side evaluation by domain experts (APUSH teachers)



Method 1: TF-IDF weighted text similarity

- For each site, concatenate all results from our reference query set
- Tokenize HTML and stem words using the Porter stemmer
- Compute TF-IDF weighted cosine distance between the textbook and this synthetic document

Method 2: Topicality using knowledge bases

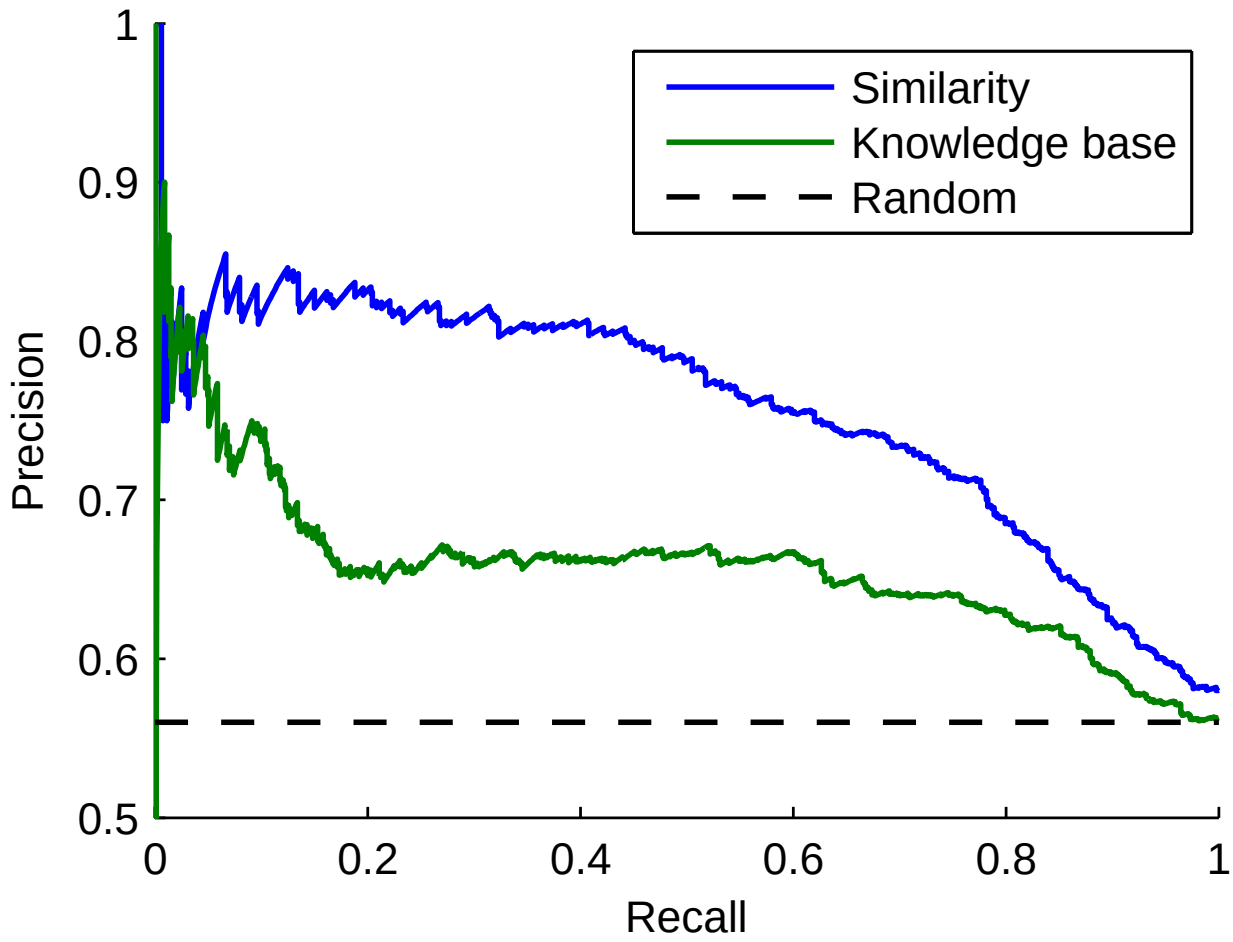
- Map proper nouns to DBpedia entities using search ("Abe Lincoln" $\rightarrow Abraham\_Lincoln$)
- DBpedia entities have categories ($Abraham\_Lincoln \rightarrow American\_Presidents$, $Illinois\_Lawyers$, $Assassinated\_HeadsOfState$)
- Form

$$CategoryScore = \frac{\#Textbook}{\#DBpedia} \tag{1}$$

where $\#Textbook$, $\#DBpedia$ count number of occurences of an entity

- Rank sites according to their category scores

Evaluation of automated methods



Conclusions

- Generic search engines with 2-3 word queries cannot fully represent the educational context
- The course textbook provides authoritative text and structured data
- Future work around how to best utilize this information
- Current version available at `http://guha.com/apushcse.html`

Approximating relevance feedback

- Reuse manually curated list by randomly selecting 50 good/bad sites (in practice this would come from usage logs)

- Augment textual similarity by

$$RelTextScore = TextScore + GoodScore - BadScore \quad (2)$$

where $GoodScore$ and $BadScore$ are textual similarity between the good/bad sites

- Augment knowledge base scoring by

$$RelCategoryScore = CategoryScore + \#Good + \#Bad \quad (3)$$

where $\#Good$ and $\#Bad$ are the number of good and bad sites associated with a category.

Evaluation with relevance feedback and hybrid scoring