Teaching an Information Retrieval Course

David Kauchak
Middlebury College
dkauchak@cs.middlebury.edu
Information Retrieval (IR)

Study of searching and processing “document” collections

- web pages
- (micro) blogs
- e-mail
- databases
- audio
- image
- video
Today...

Course content overview

Why teach an IR course?

Sample assignments: building an IR system from scratch
Designing an IR course

- books
- other courses


IR Course Content

The basics
- text processing
- IR system features and uses
- Index construction
- Evaluation
- IR and the web

IR extensions
- alternate media (image, audio, video)
- query expansion
- online advertising
- cross-lingual IR
- snippet generation
- relevance feedback
- question answering
- parallel computing
- machine learning
- natural language processing
- algorithms/data structures
### Course variability

**Basics**

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Reading</th>
<th>Slides/Handouts</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/2</td>
<td>Admin. material, Introduction</td>
<td>Ch. 1 except 1.2</td>
<td>admin, slides, pdf</td>
</tr>
<tr>
<td>9/7</td>
<td>Text pre-processing</td>
<td>Ch. 2, 5.1</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>9/9</td>
<td>Index construction</td>
<td>Ch 1.2, Ch. 4</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>9/14</td>
<td>Index compression</td>
<td>Ch. 5</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>9/16</td>
<td>TF-IDF</td>
<td>Ch. 6 except 6.4.4</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>9/21</td>
<td>Faster TF-IDF</td>
<td>Ch. 7, article</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>9/23</td>
<td>Evaluation</td>
<td>Ch. 8</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>9/28</td>
<td>Spelling correction</td>
<td>Ch. 3.3, 3.4, article</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>9/30</td>
<td>Relevance feedback/query expansion</td>
<td>Ch. 9</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>10/5</td>
<td>Web search basics</td>
<td>Ch. 19 (except 19.3), article</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>10/7</td>
<td>Crawling</td>
<td>Ch. 20</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>10/12</td>
<td>Link Analysis</td>
<td>Ch. 21</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>10/14</td>
<td>Midterm</td>
<td>Ch 21</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>10/19</td>
<td>full recess</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/21</td>
<td>Text segmentation</td>
<td>paper, article</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>10/26</td>
<td>Audio processing basics</td>
<td>paper</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>10/28</td>
<td>Audio search</td>
<td>paper</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>11/2</td>
<td>Image processing basics</td>
<td>paper, article</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>11/4</td>
<td>Project proposal discussion</td>
<td>paper</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>11/9</td>
<td>Document Image search</td>
<td>paper</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>11/11</td>
<td>Information Extraction</td>
<td>paper, article</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>11/12</td>
<td>Rose Hills (substitute lecture)</td>
<td>paper</td>
<td></td>
</tr>
<tr>
<td>11/16</td>
<td>Text classification</td>
<td>Ch. 13 (except 13.5), 14.intro, 14.1, 14.3-6, 15-15.3</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>11/18</td>
<td>Text classification2</td>
<td>article</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>11/23</td>
<td>Text clustering</td>
<td>Ch. 16</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>11/25</td>
<td>No class, substituted on 11/12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/30</td>
<td>Hierarchical clustering</td>
<td>Ch. 17, paper, article</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>12/2</td>
<td>Online Advertising</td>
<td>Ch. 3.9.2, 19.3</td>
<td>slides, pdf</td>
</tr>
<tr>
<td>12/7</td>
<td>Ethics in IR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12/9</td>
<td>Review (cross-lingual IR?)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12/14</td>
<td>Final time 9am - project presentations</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Extensions**

http://www.cs.pomona.edu/~dkauchak/classes/f09/cs160-f09/
Course variability

basics

1. Evaluation
2. Retrieval models
   - Language modeling, Boolean
   - Vector space, Latent Semantic Indexing
   - Probabilistic IR
3. Statistics of text
4. Indexing models (storing and accessing)
5. File organization
6. Efficiency, possibly including compression
7. Clustering
8. Relevance feedback

extensions

http://ciir.cs.umass.edu/cmps ci646/syllabus.html
# Course Variability

**Basics**

<table>
<thead>
<tr>
<th>Session</th>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jan 26</td>
<td>Cancelled!</td>
</tr>
<tr>
<td>2</td>
<td>Feb 2</td>
<td>Structure of IR Systems</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Evidence from Terms</td>
</tr>
<tr>
<td>3</td>
<td>Feb 9</td>
<td>Vector Space Models</td>
</tr>
<tr>
<td>4</td>
<td>Feb 16</td>
<td>Language Models</td>
</tr>
<tr>
<td>5</td>
<td>Feb 23</td>
<td>Evidence from Behavior</td>
</tr>
<tr>
<td>6</td>
<td>Mar 1</td>
<td>Evidence from Metadata</td>
</tr>
<tr>
<td>7</td>
<td>Mar 8</td>
<td>User Interaction</td>
</tr>
<tr>
<td>8</td>
<td>Mar 15</td>
<td>Evaluation</td>
</tr>
<tr>
<td>9</td>
<td>Mar 22</td>
<td>Spring Break</td>
</tr>
<tr>
<td>10</td>
<td>Apr 5</td>
<td>Cross-Language Retrieval</td>
</tr>
<tr>
<td>11</td>
<td>Apr 12</td>
<td>Document Image Retrieval</td>
</tr>
<tr>
<td>12</td>
<td>Apr 19</td>
<td>Speech and Music retrieval</td>
</tr>
<tr>
<td>13</td>
<td>Apr 26</td>
<td>Photograph and Video Retrieval</td>
</tr>
<tr>
<td>14</td>
<td>May 3</td>
<td>Project Reports</td>
</tr>
<tr>
<td>15</td>
<td>May 10</td>
<td>Final Exam</td>
</tr>
</tbody>
</table>

**Extensions**

http://terpconnect.umd.edu/~oard/teaching/796/spring04/syllabus.html
Why IR?

Relevant application
- IR systems show up everywhere!
- Good skills for finding a job

Low barrier to entry
- Technically I only require CS2 (though some further maturity helps)

Good combination of NLP and data structures

Flexible course design
- Bias towards NLP
- Bias towards alternate media
- Bias towards cluster computing
- Bias towards theory
- Bias towards software development
- …
Sample IR assignments

Develop a working IR system from the ground up

High-level goals

- Reinforce the classroom concepts
- Improve programming skills and experience more realistic software project environment
- Introduce students to the research process
Project overview

4 structured assignments
- cover the basic IR concepts
- assignments are cumulative
- build a barebones IR system

Final project
- extend the basic system
- student selected topics
- research-oriented
- open-ended
Assignment 1: Text processing

Implement text processing methods

• tokenization
• casing
• normalization
• stemming
• lemmatization
• stop-word removal
• ...

Analyze the impact

<table>
<thead>
<tr>
<th>Modifier</th>
<th>Vocab size</th>
</tr>
</thead>
<tbody>
<tr>
<td>simple tokenization</td>
<td>198K</td>
</tr>
<tr>
<td>improved tokenization</td>
<td>114K</td>
</tr>
<tr>
<td>number folding</td>
<td>108K</td>
</tr>
<tr>
<td>lowercasing</td>
<td>95K</td>
</tr>
<tr>
<td>stemming</td>
<td>91K</td>
</tr>
<tr>
<td>stop list</td>
<td>114K</td>
</tr>
<tr>
<td>num fold+lower+top</td>
<td>89K</td>
</tr>
<tr>
<td>all</td>
<td>68K</td>
</tr>
</tbody>
</table>
Assignment 2: Boolean queries

Building an inverted index

Support boolean query operations
- AND: intersection of two entries
- OR: union of two entries
- NOT: complement of an entry

Education
- 2 \rightarrow 4 \rightarrow 8 \rightarrow 16 \rightarrow 32 \rightarrow 64 \rightarrow 128

AI
- 1 \rightarrow 2 \rightarrow 3 \rightarrow 5 \rightarrow 8 \rightarrow 13 \rightarrow 21

Banana
- 13 \rightarrow 16
Assignment 3: Ranked IR

- Query and document are represented as word count vectors
- Documents are ranked by the similarity between vectors
- Implement and explore possible variations on rankings
Assignment 4: Evaluation

- Implement different evaluation measures
  - precision
  - recall
  - mean average precision (MAP)
  - normalized discounted cumulative gain (NDCG)
  - …

- Analyze the results along many dimensions (ranking, text processing)

<table>
<thead>
<tr>
<th>Count norm.</th>
<th>Term weighting</th>
<th>Length norm.</th>
<th>Precision@20</th>
<th>Recall@20</th>
<th>RPrecision</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>none</td>
<td>none</td>
<td>0.006</td>
<td>0.052</td>
<td>0.002</td>
<td>0.015</td>
</tr>
<tr>
<td>log</td>
<td>none</td>
<td>none</td>
<td>0.029</td>
<td>0.224</td>
<td>0.060</td>
<td>0.076</td>
</tr>
<tr>
<td>none</td>
<td>IDF</td>
<td>none</td>
<td>0.046</td>
<td>0.390</td>
<td>0.109</td>
<td>0.158</td>
</tr>
<tr>
<td>none</td>
<td>none</td>
<td>cosine</td>
<td>0.037</td>
<td>0.308</td>
<td>0.100</td>
<td>0.158</td>
</tr>
<tr>
<td>log</td>
<td>IDF</td>
<td>none</td>
<td>0.055</td>
<td>0.455</td>
<td>0.132</td>
<td>0.189</td>
</tr>
<tr>
<td>none</td>
<td>IDF</td>
<td>cosine</td>
<td>0.065</td>
<td>0.539</td>
<td>0.148</td>
<td>0.218</td>
</tr>
<tr>
<td>log</td>
<td>IDF</td>
<td>cosine</td>
<td>0.068</td>
<td>0.547</td>
<td>0.166</td>
<td>0.229</td>
</tr>
</tbody>
</table>
Code base after the assignments

- 27 classes
- ~2400 lines of code
- The students have been working on this code base throughout the entire course
Final project

Student requirements

- Implement an addition to the system related to class
- Work in teams
- Integrate with current codebase
- Evaluate
- Give write-up and presentation

Goals

- Explore advanced topics in IR that interest them
- “Real” software development
- An introduction to the research process
Final project goals: explore advanced topics

- GUI development
- Query optimization
- Query suggestion/rewriting
- Index compression
- Faster/approximate ranked retrieval
- Relevance feedback
- Result clustering
- Result classification
- Web crawling

- Position index/phrasal queries
- Wildcard/regex support
- Document segmentation
- Snippet generation
- Parallelized indexing
- Parallelized querying
- Multimedia search
- Document importance (e.g. PageRank)
- Advertising
Final project goals:
real software development

Must work in teams

Must integrate with other teams

Use code repository (Git, SVN)

Final project size was over 40 classes and over 6000 lines of code
Final project goals:
explore research process

Evaluate the impact of the addition

Write-up results in a research format

Peer review research write-ups

Give short, oral presentation summarizing project and results
Experiences

Pomona College (Fall 2009)
- 13 students
- Initial four assignments required roughly 2 weeks each
  - which were done in tandem with some written work
- Implemented in Java

Middlebury College (Fall 2012)
- 11 students
- Minor tweaks, but overall same structure
Experiences

Assignments
- Reinforce good coding practices
- Regular expressions
- Actual use for implementing your own linked list!
- Experiment design and data analysis

Final projects
- 6 different groups
- Integration wasn’t too painful (though I did much of the coordination)
- Groups working on same topic (e.g. snippet generation) were able to compare results
- Evaluation was the most challenging part for most groups
• Supports full text queries with ranking
• Supports boolean queries
• Snippet generation
• 15K data set
• Works very efficiently
Clinton to American youth: don't inhale
... A stupid thing to do ... It is a stupid thing to ...

Treasury official blasts anti-crime budget cuts
... proposed cutbacks in his budget as stupid. ... This is stupid ...

Half Way Mark For GOP Scorecard Shows Points Drop
... stupid test to regulations. And if it is stupid, ... they are substitution appeals to fear for a discussion of the facts and the ...

Czech editor gets suspended sentence for racism
... Jews and half-breeds in December 1992 and started a campaign to brand the ... the stupid Czech nation merely sits back and watches this Jewish looting going ...

Monty Python's Flying Circus Celebrates 25 Years
... of Monty Python's Flying Circus would have to calm down a bit to be called ... weren't- Usually when they were doing something stupid, they also were being incredibly intelligent ...
Flexibility

Four assignments cover the basics of building a search engine

Many components can be easily altered based on the particular needs of the course

Final projects can be adapted to course particular course material

For larger classes
- interfaces well defined: grading scripts are very useful
- may have students to do less written work
- split final projects into different IR systems
EAAI

Educational Advances in Artificial Intelligence

http://eaai.cs.mtu.edu/
Resources

http://www.cs.middlebury.edu/~dkauchak/ir_project/

- Two variants of the course:
- Course syllabus with slides
- 4 assignment descriptions (pdf and .tex)
- Final project description (pdf and .tex)
- Starter AND solution code in Java
  - Contact me for these
- Sample final project
  - demo and student final papers
- Sample grading scripts (though these should be considered in beta 😊)