http://www.flickr.com/photos/56685562@N00/565216/

Document Image Retrieval

David Kauchak
cs458
Fall 2012

adapted from:
David Doermann
http://terpconnect.umd.edu/~oard/teaching/796/spring04/slides/11/796s0411.ppt

Admin

Schedule
- Friday by 6pm Assignment 4 writeup
- Sunday: Sent me an e-mail with team and topic
- Tuesday:
  - Quiz!
  - Project proposal draft
  - Project proposal discussion
- Thursday: Finalized project proposal

Rest of the semester…

Quiz 2?

Grading

Assign 4 write-ups

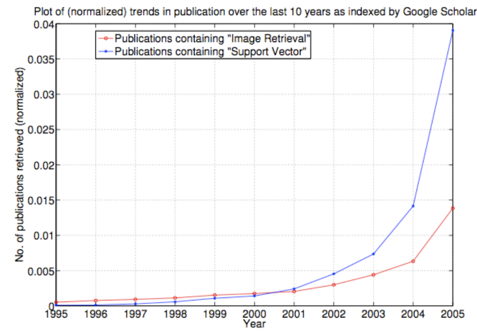
- Some general comments
  - explain data set and characteristics
  - explain your evaluation measure(s)
  - think about the points you're trying to make, then use the data to make that point
  - comment on anything abnormal or surprising in the data
  - dig deeper if you need to
  - if you have multiple evaluation measures, use them to explain/ understand different behavior
  - try and explain why you got the results you obtained

Information retrieval systems

Spend 10 minutes playing with three different image retrieval systems

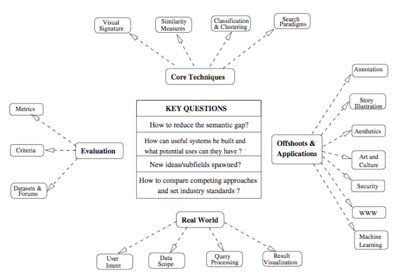
- http://en.wikipedia.org/wiki/Image_retrieval has a number
- What works well?
- What doesn't work well?
- Anything interesting you noticed?

Image Retrieval



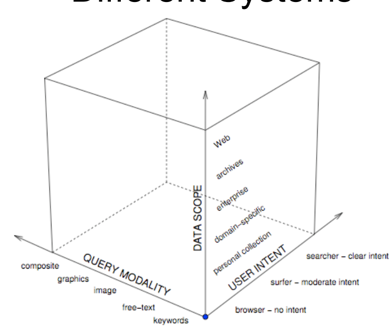
http://infolab.stanford.edu/~wangz/project/imsearch/review/JOUR/datta_TR.pdf

Image Retrieval Problems



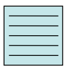
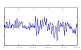
http://infolab.stanford.edu/~wangz/project/imsearch/review/JOUR/datta_TR.pdf

Different Systems

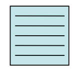
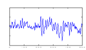


http://infolab.stanford.edu/~wangz/project/imsearch/review/JOUR/datta_TR.pdf

Information retrieval: data

	amount of data	data characteristics
Text retrieval 	trillions of web pages within an order of magnitude in "private" data	• user generated • some semi-structured • link structure
Audio retrieval 	order of a few billion? last fm has 150M songs	• mostly professionally generated • co-occurrence statistics
Image retrieval 	This is blowing up! - 60 photos/sec uploaded via instagram - 4.5 million photos/day Flickr - > 100 billion photos on Facebook	• user generated • becoming more prevelant • some tagging • incorporated into web pages (context)

Information retrieval: challenges

	challenges	other dimensions?
Text retrieval 	• scale • ambiguity of language • link structure • spam	
Audio retrieval 	• query language • user interface • features/pre-processing	
Image retrieval 	• query language • user interface • features/pre-processing • ambiguity of pictures • data size	

Today: Document Image Search

Why not general image search?

What's in a document?

- I give you a file I downloaded
- You know it has text in it
- What are the challenges in determining what characters are in the document?
  - File format:

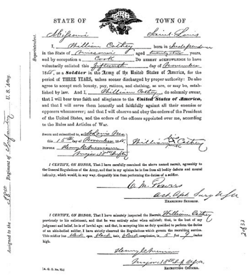
1. What file types are returned in a Google search?

There are 13 main file types searched by Google in addition to standard web formatted Microsoft Office formats:

- Adobe Portable Document Format (pdf)
- Adobe PostScript (ps)
- Lotus 1-2-3 (wk1, wk2, wk3, wk4, wk5, wki, wks, wku)
- Lotus WordPro (lwp)
- MacWrite (mw)
- Microsoft Excel (xls)
- Microsoft PowerPoint (ppt)
- Microsoft Word (doc)
- Microsoft Works (wks, wps, wdb)
- Microsoft Write (wri)
- Rich Text Format (rtf)
- Shockwave Flash (swf)
- Text (ans, txt)

http://www.google.com/help/faq_filetypes.html

Is this a document?




The Washington Post

After Zarqawi, No Clear Path In Weary Iraq

The Young Apprentice

How U.S. Forces Found Iraq's Most-Wanted Man

Kaine Delays Execution Of Inmate for 6 Months

Document Images

A document image is a document that is represented as an image, rather than some predefined format

Like normal images, contain pixels
- often binary-valued (black, white)
- But greyscale or color sometimes

300 dots per inch (dpi) gives the best results
- But images are quite large (1 MB per page)
- Faxes are normally 72 dpi

Usually stored in TIFF or PDF format

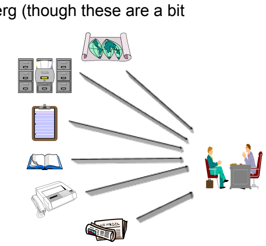
Want to be able to process them like text files

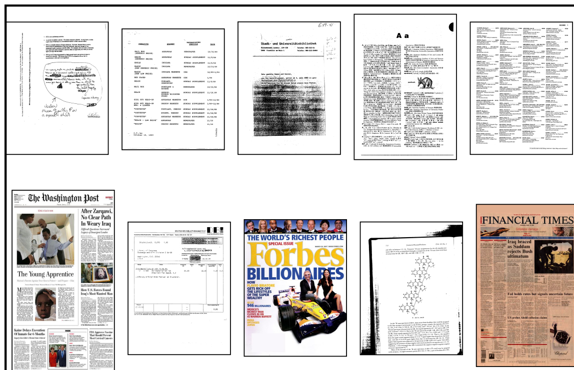
Sources of document images

Web
- http://dli.iiit.ac.in/
- Arabic news stories are often GIF images
- Google Books, Project Gutenberg (though these are a bit different)

Library archives

Other
- Tobacco Litigation Documents
  - 49 million page images



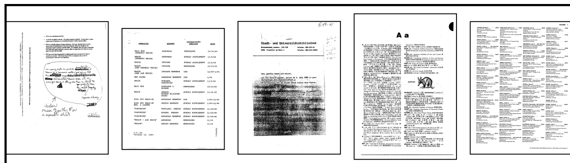


Document Image Database

Collection of scanned images

Need to be available for indexing and retrieval, abstracting, routing, editing, dissemination, interpretation

NOTE: more needs than just searching!



What are the challenges?
What are the sub-problems?

Challenges

They're an image ☺

Quality
- scan orientation
- noise
- contrast

Hand-written text

Hand-written diagrams

Additional Reading

- A. Balasubramanian, et al. Retrieval from Document Image Collections, *Document Analysis Systems VII*, pages 1-12, 2006.
- D. Doermann. The Indexing and Retrieval of Document Images: A Survey. *Computer Vision and Image Understanding*, 70(3), pages 287-298, 1998.