Brad Burkman's Notes for

CSCE 561 Data Storage and Retrieval
Dr. Aminal Islam
Fall 2018

Contents

1 Monday 20 August: Introduction

Slides from first day are on Moodle.

Professor Dr. Aminal Islam

- Natural Language Processing (NLP)
- Data Mining
- Machine Learning
- Artificial Intellgence
- Data Analytics
- Word Segmentation (2007)
- Textual Error Correction

Steve Jobs: "Creativity is just connecting things."

When something doesn't work in your research, you're on the right track.

Final product of course will be a research paper

- Pick two articles from ACM SIGIR (Special Interest Group on Information Retrieval) 2018
- Write a review of each, and present.
- Find a problem that leads to an original discovery.
- Write it up.

Text: *Introduction to Information Retrieval* by Christopher D. Manning et al.
`https://nlp.stanford.edu/IR-book/`

Classmates

2 Notes from Skimming SIGIR'18 Abstracts

Sessions

Keynotes
1A: New IR Applications
1B: Log Analysis
1C: Prediction
1D: Learning to Rank I
2A: Sentiment and Opinion
2B: Social
2C: App Search and Recommendation
2D: Conversational Systems
3A: Social Good
3B: Privacy
3C: Question Answering
3D: Learning to Rank II
4A: Fairness and Robustness
4B: Behavior
4C: Medical and Legal IR
4D: Recommender Systems - Methods
5A: Location and Trajectory
5B: Entities
5C: New Metrics
5D: Recommender Systems - Applications
6A: Evaluation
6B:: Hashing and Embedding
6C: Knowledge Bases/Graphs
6D: Mobile User Behavior
7A: Crowdsourcing and Assessment
7B: Content and Semantics
7C: Interfaces
Short Research Papers I
Short Research Papers II
Demonstration Papers I
Demonstration Papers II
Sirip: Industry Days

Tutorials
Workshops
Doctoral Consortium

Knowledge Graphs (KG)
Knowledge Bases/Graphs
The potential parent-child relationships linking the new concepts to the existing ones are then predicted using a set of semantic and graph features.

3 Wednesday 22 August: Information Retrieval (IR)

Jesse and Nusrat
Shekufeh

Is an image a document? Not the image itself, but the metadata is.

Comparing the query text to the document text and determining what is a good match is the *core issue* of IR.

Google Trigram Model for Relatedness `ares.research.cs.dal.ca/gtm/`

SemEval annual competition since 2012. International Workshop on Semantic Evaluation, related to the Association for Computational Linguistics.

Three components of a search algorithm:
Document Representation
Query Representation
Retrieval Model or Ranking Model.

"corpus" (singluar) "corpora" (plural)

Concerns in IR: Relevance and Efficiency

Relevance
Optimized based on location
Proper subject
Timely
Authoritative, based on other sites linking to it.
Satisfying goals of the user.

"Bag of Words": Frequency count, no word order.

Intelligent IR
Takes into account the meaning of the words used.
Order of words
Indirect feedback
Trustworthiness

IR is not just web search.

4 Monday 27 August: History and Dimensions of IR

Reviewing Wednesday:

- Main concerns in IR: Relevance and Efficiency
- Three components of a search algorithm:
  - Document Representation
  - Query Representation
  - Ranking Model (Retrieval Model)

Simplicity is more important than relevance or efficiency.
Google front page is getting simpler over time.

Dimensions of IR
Different media, types of search applications, tasks

- Video, Photos, Music, Speech
- Like text, content is difficult to describe and compare.

Recommendation Systems (Amazon, Netflix)
Question Answering

- Information Extraction Problem

Text Mining

- Topic Modeling
- "Stock words" v/s "functional words"

- Text Clustering: No labels, Unsupervised learning
- Text Categorization/Classification: Labels, Machine learning
- NER, Named Entity Recognition

Automated Document Categorization
Information Filtering (Spam Filtering)
Automated Document Clustering
Information Integration

Database Schema Mapping for merging databases

At the end of the course, we'll talk about data mining methods.

Summarization can be *Extractive* or *Abstractive*

Extractive	Find n most important sentences.
	Order the sentences based on their importance.
	Don't change the sentences.
Abstractive	Change, merge sentences. Summarize. *Natural Language Generation (NLG)*

Text Mining / Text Analytics

History of IR

1960’s - 1970’s	Text Retrieval Systems	
	Law and Medicine	
		Finding precedents
		Many law firms have their own proprietary search systems
	Boolean and Vector-space Models	
	Professor Salton at Cornell	
1980’s	Large document database systems	
		Lexis-Nexis
		Dialog
		MEDLINE / PubMed
1990’s	Searching FTP’able documents	
		Archie
		WAIS
	Searching WWW	
		Lycos
		Yahoo
		AltaVista
	Organized Competition	
		NIST TREC
	Recommender Systems	
		Ringo
		Amazon
		Net Perceptions
	Automated Text Categorization and Clustering	
2000’s	Link Analysis for Web Search (Google)	
	Automated Information Extraction	
	Parallel Processing (MapReduce)	
	Question Answering (TREC Q/A Track)	
	Multimedia IR	
	Cross-Language IR	
		DARPA Tides was a a failure because the translation algorithms were poor
	Document Summarization	
	Learning to Rank	
2010’s	Intelligent Personal Assistants	
	Complex Question Answering (IBM Watson)	
	Deep Learning (Neural Networks)	
	Distributional Semantics (Dr. Aminal’s research)	
		Summarizing raw text (?)

Choosing a Paper to Read
New Application, New Method, or New Problem?
New Application is a Master's Thesis requirement
New Problem is best, even if the solution is very naive.
"Add a statement in the field of knowledge that was previously unknown."

Anecdotes

Online advertising last year, $83B USD
Surpassed cable TV revenue last year.

Most popular Google search query as of 2015, and motivator for the creation of Google Image Search.
"Jennifer Lopez's Green Dress."

Why did Google beat Yahoo?
Connected things.
Incorporated user feedback.

Questions to Alexa are usually *very* frequently asked questions. Alexa doesn't even send the speech to Apple for parsing and results. It already has them in memory. As Alexa has more experience with the user, it becomes even more **pseudoefficient.**

5 Wednesday 29 August: