COMP 4321 Search Engines for Web and Enterprise Data

Course Outline

Week	Topics	Chapters	Slides	Supp Slides	Notes	Ref Materials
1	Introduction and course overview					
	○ IR vs. DBMS; search engine applications	[MRS] Ch 1 [BR] Ch 1	slides	"Take Home Message" for 000-se-intro.ppt **new!**	Notes	
	○ Search engine business models and industry	[MRS] Ch 19	slides			• Regulating Google as a Utility **new!** • Albert Yeung against Google on query suggestions **new!** Example **new!** • http://www.wolframalpha.com • Interview of Wolfram (9/23/2010)
2-4	Information retrieval models and Inverted Files					
	○ Boolean model, document ranking, vector space model	[MRS] Ch 1, 6 [V] Ch 5	slides	-	-	• Historical work on term weighting and similarity **new!**
	○ Inverted files and extension; fast main-memory-based indexing	[MRS] Ch1 [V] Ch 4	slides			
	○ Extended Boolean model	[MRS] Ch 1	slides (for discussion only)		Notes	
5-6	Web-based information retrieval					
	○ Hypursuit, WISE, Google, PageRank, Clever	[MRS] Ch 19, 20 [BR] Ch 13	slides Clever	-	-	• How Baidu Uses Deep Learning to Drive Success on the Web (9/2014) **new!** • Hypursuit **new!** • WISE algorithm **new!** • WISE system **new!** • Nature article (1999) • Size estimation of the web **(slides)** • Commentary on Google and Kleinberg's work • Google's alleged 200 parameters
6-7	Pattern matching					
	○ Brute-force method		slides	-	-	
	○ Knuth-Morris-Pratt (KMP) method					Boyer and Moore
	○ Regular expressions and finite state automata	[MRS] Ch 12				Online FSA demo (English version) This Dutch version works
7-8	Retrieval effectiveness and performance measures					
	○ Precision, recall and fallout; standard document	[MRS] Ch 8 [V] Ch 7	slides	-	-	Google Stats **new!** Google VP Engineering on search quality (2008) **new!** Google's new quality factors:

	collections for benchmarking					expertise, authority and trust (2014) **new!**
8-10	Document preprocessing					
	○ Stopword removal and Stemming	[MRS] Ch 2 [V] Ch 2 [BR] appendix [FB] Ch 8	slides		Notes	
	○ Index term selection and term discrimination values	[MRS] Ch 5 [V] Ch 2	slides			
	○ Thesauri, term phrase formation	[MRS] Ch 9 [V] Ch 2 [FB] Ch 5	slides			
	○ Collocation analysis for phrase extraction		slides			
	○ Thesaurus, taxonomy, ontology, Wordnet	[MRS] Ch 9 [V] Ch 2	slides	Old+new slides in one set **new!** Only the new slides **new!**		Metaweb video Apture video
11	Relevance feedback and functionality overview					
	○ Implicit vs. explicit feedback	[MRS] Ch 9 [V] Ch 5	slides (shortened)			• What Do People Want from Information Retrieval? [the paper is not needed in final exam] • An overview paper on ranking and relevance feedback [the paper is not needed in final exam but the slides are included.]
	○ Search engine personalization		slides			
12	Enterprise search					
	○ Differences from Web search; issues and challenges	[MRS] Ch 4 (only a little)	slides			○Enterprise Search: Tough Stuff. Queue, Apr 2004 [the paper is not needed in final exam]
12	Clustering, signature files and course summary					
	○ In-class project demonstration					
	○ Signature files and superimposed coding	[FB] Ch 4, [BR] Ch 8.3	slides (shortened)			
	○ Course summary		slides			