UNIVERSITY OF NEBRASKA AT OMAHA
COURSE SYLLABUS/DESCRIPTION

Department and Course Number	CSCI 3850
Course Title	Foundations of Web Search Technologies
Course Coordinator	Parvathi Chundi
Total Credits	3
Date of Last Revision	February 25, 2015

1.0 Course Description Information:

1.1 Catalog description:

This course provides students a basic understanding of how search and information flow works on the web. Main topics include: document representation, inverted indexing, ranking of web search results, vector-space model, web graph, PageRank, search-based advertising, information cascades, and web crawling.

1.2 Prerequisites of the course:

CSCI 2030
CSCI2850
Instructor Approval

1.3 Overview of content and purpose of the course:

In this course, students will learn the theoretical concepts, algorithms, and strategies used in building web search engines, commonly used measures such as precision and recall to evaluate web search engines, the Web graph and Page rank computation, advertising on the internet, mathematical models of information cascades, crawling the Web, and the ethics of web search.

1.4 Unusual circumstances of the course.

None.

2.0 Course Justification Information

2.1 Anticipated audience / demand:

All students are in the IS&T college are able to take the course.

The course is not a service-learning course

Students from other colleges are not required to take this course.

2.2 Indicate how often this course will be offered and the anticipated enrollment:

The course will be offered every Fall semester.

It is expected that 15-30 students would take the course.

2.3 If it is a significant change to an existing course, please explain why it is needed:

This is a new course.

3.0 List of performance objectives stated in learning outcomes in a student’s perspective:

3.1 Understand the need for web search and ranking systems
3.2 Understand the content representation for web search
3.3 Understand efficient ways to index the web content
3.4 Understand the vector space model
3.5 Learn how to evaluate web search systems
3.6 Understand how Web is represented as a graph
3.7 Understand the concept of PageRank and its computation
3.8 Learn the information cascade models on the Web
3.9 Learn how search and web advertising are correlated
3.10 Understand the concepts behind web crawling
3.11 Ethical implications of web search and crawling

4.0 Content and Organization

4.1 Introduction ` (1.5 hours)
4.1.1 Web Search Architecture
4.1.2 Structure of the Web Graph
4.1.3 Retrieval Vs Ad Hoc Filtering
4.1.4 Ranking
4.2 Classic Information Retrieval (1.5 hours)
4.2.1 Information Vs Data Retrieval
4.2.2 Logical View of Documents
4.2.3 Boolean Model
4.3 Pre-processing and Indexing (6 hours)
4.3.1 Tokenization, Normalization, Stemming
4.3.2 Posting Lists and Skip Lists
4.3.3 Bi-Word and Positional Indexes
4.3.4 Inverted Index Construction
4.4 Vector Space Model (6 hours)
4.4.1 Term Weighting
4.4.2 Processing of Web Searches and Ranking
4.4.3 Incorporating the Index
4.5 Evaluating Web Search (3 hours)
4.5.1 Assessing Relevance
4.5.2 Precision and Recall
4.5.3 User Utility
4.6 The Structure of the Web (3 hours)
4.6.1 The Web As a Directed Graph
4.6.2 Information Networks and Hypertext
4.6.3 Bow-Tie Structure
4.7 Link Analysis and Web Search (7.5 hours)
4.7.1 Hubs and Authorities
4.7.2 PageRank
4.7.3 Link Analysis in Web Search
4.8 Sponsored Search Markets (7.5 hours)
4.8.1 Advertising Tied to Search Behavior

4.8.2 Advertising as Matching Market

4.8.3 The VCG Principle and Analysis

4.9 Information Cascades (6 hours)

4.9.1 Following Crowds

4.9.2 A Simple Herding Experiment

4.9.3 Decision Making Under Uncertainty

4.9.4 A Simple, General Cascade Model

4.9.5 Sequential Decision Making and Cascades

4.10 Web Crawling (3 hours)

4.10.1 Basic Crawlers

4.10.2 Implementation Issues

4.10.3 Universal Crawlers

4.10.4 Topical Crawlers

4.11 Search Engines and Ethics (3 hours)

3.11.1 Search Engine Bias and Problem of Opacity

3.11.2 Privacy, Consent, and Non-voluntary Disclosure

3.11.3 Monitoring and Surveillance

5.0 Teaching Methodology

5.1 Methods to be used.

The basic teaching method will be the instructor’s lectures.

5.2 Student role in the course.

The students will attend lectures, complete assignments, and pass all exams. Students will also participate in discussions and problem solving sessions held during class.

5.3 Contact hours.

Three (3) hours a week.

6.0 Evaluation

Students will be evaluated on their performance in exams, home works, and classroom participation. The material for home works, exams, and class room discussions will be designed by the instructor using the examples and end-of-chapter exercises from the recommended text books and other reading materials. Basis of evaluation includes understanding of basic concepts in the fundamentals underlying the internet search technologies and information networks including the inverted indices, skip lists, vector space model, precision and recall, web graph, PageRank, information cascades, and web crawling. Final grade will be derived as –

Exams (2 midterms):	20%
Home works	50%
Classroom participation/discussion	10%
Final exam:	20 %

Grading scale and criteria.

Percent	**Grade**	**Percent**	**Grade**
97 – 100	A+	77 – 79	C+
94 – 96	A	70 – 76	C
90 – 93	A–	70 – 73	C–
87 – 89	B+	67 – 69	D+
84 – 86	B	64 – 66	D
80 – 83	B–	60 – 63	D–
		0—59	F

7.0 Resource Material

The material from the following three text books will be distributed to the students via Blackboard.

7.1 Required Text Books

7.1.1 Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008.

7.1.2 David Easley and Jon Klienberg, *Networks, Crowds, and Markets Reasoning about a Highly Connected World*, Cambridge University Press, 2010.

7.1.3 Christopher D. Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, The MIT Press 1999.

7.2 Other suggested reading materials, if any.

7.2.1 T. Blanke,Ethical Subjectification and Search Engines: Ethics Reconsidered, *International Review of Information Ethics*, 3: 33–38, 2005.

7.2.2 B. Friedman and H. Nissenbaum, Bias in Computer Systems, *ACM Transactions on Computer Systems*, 14(3): 330–347, 1996

7.2.3 E. Goldman, Search Engine Bias and the Demise of Search Engine Utopianism. In *Web Search: Multidisciplinary Perspectives*. Eds. A. Spink and M. Zimmer, Berlin: Springer-Verlag, pp. 121–134, 2008.

7.2.4 K. E. Himma, Privacy vs. Security: Why Privacy is Not an Absolute Value or Right, *University of San Diego Law Review* (Fourth Annual Editors' Symposium), 45: 857–921, 2007.

7.3 Current bibliography and other resources.

7.3.1 W. Bruce Croft, Donald Metzler, and Trevor Strohman, Search Engines: Information Retrieval in Practice, Addison-Wesley 2011.

7.3.2 Naser El-Bathy and Ghassan Azar, Intelligent Information Retrieval and Web Mining, Lambert Academic Publishing, 2010.

7.3.3 Bing Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Springer 2011.

7.3.4 Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley 2011.

7.3.5 Robert R. Korfhage, Information Storage and Retrieval, Wiley 1999.

7.3.6 Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, Morgan Kauffman, 2011.

7.3.7 Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein, Introduction to Algorithms, The MIT Press, 2011.

7.3.8 David A. Grossman and Ophir Frieder, Information Retrieval, Springer, 2004.

7.3.9 Robert A. Hanneman and Mark Riddle, Introduction to Social Network Methods, 2005.

7.3.10 Stanley Wasserman and Katherine Faust, Social Network Analysis: Methods and Applications (Structural Analysis in Social Sciences), Cambridge University Press, 1994.

7.4 Other sources of information.
Apache Lucene Project

8.0 Other Information:

8.1 Accommodations statement:

Reasonable accommodations are provided for students who are registered with Disability Services and make their requests sufficiently in advance. For more information, contact Disability Services (MBSC 111, Phone: 554-2872, TTY: 554-3799) or go to the website: www.unomaha.edu/disability.

8.2 Other:

8.3 Author(s): Parvathi Chundi

9.0 (Fill out for CSCI and CIST courses) Estimate Computer Science Accreditation Board (CSAB) Category Content (class time in hours):

CSAB Category	Core	Advanced
Data structures		30
Computer organization and architecture		
Algorithms and software design		15
Concepts of programming languages		

9.0 Oral and Written Communications:

Every student is required to submit at least __0__ written reports (not including exams, tests, quizzes, or commented programs) to typically ____0_ pages and to make __0__ oral presentations of typically _0__ minutes duration. Include only material that is graded for grammar, spelling, style, and so forth, as well as for technical content, completeness, and accuracy.

10.0 **Social and Ethical Issues:**
Please list the topics that address the social and ethical implications of computing covered in all course sections. Estimate the class time spent on each topic. In what ways are the students in this course graded on their understanding of these topics (e.g. test questions, essays, oral presentations, and so forth?).

a) Search Engine Bias and Problem of Opacity 1 contact hour
b) Privacy, Consent, and Non-voluntary Disclosure 1 contact hour

c) Monitoring and Surveillance 1 contact hour

Students will be graded on their responses to essay questions related to these issues included in home works and exams.

11.0 Theoretical content:

The theoretical material covered in the course includes the graph theory, probability theory, counting principles, auction principles, and mathematical models of information flow networks. These topics are discussed throughout the course.

12.0 Problem analysis:

There will be analysis on quizzes, written/programming assignments and exams. Analysis may be given during the lecturing or in the written form to be distributed in the class. Students are asked to review the analysis materials, and are responsible for these materials.

13.0 Solution design:

Students will gain a fundamental understanding of web search, ranking, information cascades, and web crawling through classroom quizzes, written assignments and exams.

CHANGE HISTORY

Date	***Change***	***By whom***	***Comments***
2/21/2014	First Version	Chundi	
4/21/2014	Second Version	Chundi	Updated references. Fixed the course description (1.0). Fixed the solution design (12.0).
2/18/2015		Chundi	Minor fixes
2/23/2015	Third Version	Chundi	Added a module on Ethics and changed the format of the document to ABET style.