AC 2012-3123: SURVEY AND ANALYSIS OF COURSES ON THE SUBJECT OF INFORMATION RETRIEVAL AND WEB SEARCH

Dr. Xiannong Meng, Bucknell University

Xiannong Meng is a professor of computer science in the Department of Computer Science at Bucknell University in Lewisburg, Penn., USA. His research interests include distributed computing, data mining, intelligent web search, operating systems, and computer networks. He received his Ph.D. in computer science from Worcester Polytechnic Institute in Worcester, Mass., USA.

Dr. Song Xing, California State University, Los Angeles

Song Xing received his B.S. and M.S. degrees in electrical engineering from Southeast University, China, in 1985 and 1990, respectively, and his Ph.D. degree in electrical and computer engineering from George Mason University, Va., in 2003. From 1985 to 1995, he was a lecturer in the Radio Engineering Department at Southeast University, China. He was also a researcher at the National Mobile Communications Research Laboratory, China, from 1990 to 1995. He was a Visiting Researcher in the Electrical and Computer Engineering departments at the University of Michigan, Dearborn, from Feb. 1995 to April 1995 and at Boston University from May 1995 to Aug. 1996, respectively. From August 1997 to May 2003, he was an instructor with the Electrical and Computer Engineering Department and a Teaching/Research assistant in the Electrical and Computer Engineering, Computer Science, and Information Systems departments at George Mason University. In 2003, Xing joined California State University, Los Angeles, where he is currently an Associate Professor in the Information Systems Department.

Ms. Wang Wei, Southeast University

Wang Wei is a CCF member and Associate Professor in the School of Computer Science and Engineering. Born in the city of ChangChun, China, Wei received a M.S. in Southeast University (SEU) in 2011. Wei worked at Southeast University since 1991, and beginning in 2011, worked at the School of Computer Science and Engineering, SEU.

Survey and Analysis of Courses on the Subject of Information Retrieval and Web Search

Abstract

In this paper, we present the result of our survey and analysis of courses on the subject of information retrieval and web search offered by colleges and universities in the U.S.A. and in China. We collected the information available on the web about 40 relevant courses. We concentrate on the goals of the course, main contents covered, textbooks used, and the student programming projects.

Most these courses are offered either for upper level undergraduate through entry level graduate students in the area of computer science and computer engineering or for students from the information science and linguistics science. These courses concentrate on the technical aspects of information retrieval. For example, the computer science courses mostly deal with information processing and search over the web while a linguistic or information science course may discuss more the language aspect and information science aspect. There is also a segment of courses designed for general audience who are the end users of information retrieval systems. These courses discuss more on the user interface, human factors, and social impact of the web search technologies.

We believe the information presented in the paper will help design, implementation, or revision of courses on the subject of information retrieval with various target audience in mind, especially for students at the undergraduate level.

1. Introduction

The pervasive nature of World Wide Web (the web) and social networks such as Facebook and LinkedIn presents tremendous opportunity and challenge for general users who use these systems everyday as well as scientists and engineers who design and build the infrastructures for these systems. One of the critical features of the web and social networks is text-based search, whether done explicitly by using search engines such as Google, or done implicitly by pressing a search button on any of these sites. The core of text-based search is supported by the theories and practices of an academic branch in computer science or information science called *information retrieval* (IR). Because of this, interests in and demands on teaching courses that deal with the subject of IR, in particular, IR in the context of the web search have risen quickly in recent years. The target audiences, thus related to the main contents, of such courses come from two academic disciplines, those who are from the library science, and those who are from computer science and software engineering. In this paper, we present the result of our survey and analysis of courses on the subject of information retrieval and web search offered by colleges and universities across the web, mainly in the United States and in China. We collected the information available on the web from about 40 courses from different colleges and universities, six of which are from Chinese universities and research institutes. We concentrate on the goals of the course, main contents covered, textbooks used, and the student projects. We believe the information presented in the paper can help design, implementation, or revision of courses on the subject of information

retrieval with various target audience groups in mind, especially for students at the undergraduate level.

The data presented here about these courses are collected through their websites. Thus the information is inherently incomplete for many reasons, such as that the instructors didn't have time to update the web content, or that some course contents are available only to internal audience through courseware such as BlackBoard or Moodle. We do our best to summarize the collected data into a coherent segment of information. The raw data used in this paper are listed in Appendix A (course websites), Appendix B (textbooks used), and Appendix C (list of course objectives and goals by each course when available). Readers can also visit the information online at one of authors' website at http://www.eg.bucknell.edu/~xmeng/webir-resources-asee2012.html.

The rest of the paper is organized as follows. Section 2 is a review of other surveys of similar nature and general discussions of teaching and learning on the subject of information retrieval and web search. In Section 3, we present our method of study and data collection. The course titles and objectives, major topics, textbooks and projects are discussed in Section 4, 5, 6, and 7, respectively, followed by some concluding remarks in Section 8.

2. A Brief Review of the Literature

We review in this section the literature concerning the teaching and learning of information retrieval. With increased interest in and importance of information retrieval and web search, more and more research projects have been on the subject of teaching and learning of information retrieval. Fenandez-Luna *et al.*[12] presented a comprehensive review of the state of teaching and learning of information retrieval. In their paper, the authors presented taxonomy, educational goals, teaching and learning methods, assessment, and curricula regarding the subject of teaching and learning of information retrieval. The authors collected and analyzed 159 papers during the 40-year period of 1968 to 2008 that has anything to do with teaching and learning of information retrieval. The authors found that about 85 percent of these papers are from the field of library science and computer science. This collection of papers gives a trend in the evolution of information retrieval. While the authors didn't give a full list of the 159 papers, they did list 104 references in their survey paper, which is invaluable to the community of information retrieval education. The British Computer Society (BCS) held two international workshops on the teaching and learning of information retrieval (TLIR 2007[6] and TLIR 2008[7]). The proceedings of these two workshops contain rich collection of papers on various subjects related to the teaching and learning of information retrieval. In the 2007 workshop, a total of 10 papers were presented, topics ranging from learning environment (e.g., E-learning), to teaching strategies (math, IR, and web search), and to curricula and evaluation. In the 2008 workshop, a total of four papers were presented. The four papers discussed the topics of teaching IR as a philosophy problem, relation between search and engines, a holistic approach to teaching IR, and a report of developing a search engine as a practical project in teaching IR. In addition, other papers have been published on the subject of teaching IR in recent years. McCown[18] contrasted his experience in teaching an IR course in which students develop a search engine from scratch with the one in which students revise code in an existing search engine. Each of the two approaches has its pros and cons, developing a search engine from scratch gives students a

greater understanding of what is behind the scene in a search engine but the end-product may be less polished, while revising existing search engine code may accomplish more functionality but students would have to overcome some steep learning curve. Zhu and Tang[33] proposed a module-based integration of IR topics into different courses in an undergraduate curriculum. Meng[20,21] presented the two cases teaching IR, one for computer science students who developed a search engine from scratch, the other for non-technical students who learned how to work with search engines and the societal impact as a result of the pervasive use of search engines.

Judging by the amount of publications and the number of courses available on the web, one can tell that overall interest in teaching and learning information retrieval in colleges and universities has been on the rise. While many aspects of teaching and learning information retrieval have been discussed in previous papers, we believe our unique contribution in this paper is to provide a survey of course contents, goals, textbooks used, and projects that are available on the web. Instructors who are interested in teaching such a course will find this collection of information useful in helping the development of a new course or revision of an existing course. Students who wish to learn the content on their own can also benefit from this collection of information.

3. Method of Study and Data Collection

The authors searched through the web for courses on the subject of information retrieval, web search, and web data mining. Each of these sites were manually visited, a few sites that didn't contain any technical content (e.g., websites that only listed a course title without any further information available on the web) were removed. We kept the sites in this survey that at the least we could identify the course title, instructor(s) of the course, and a list of main topics of the course. Most of the sites contain much richer contents than the aforementioned minimum amount of information. Among the additional information found on these sites include teaching schedules, topics discussed in the course, lecture notes, detailed homework and project assignments, and any combinations of the above. In the end, information from a total of 38 course websites is collected. Twenty-eight courses of the 38 total have been taught since 2007 (a five-year window through spring 2012). A list of these 28 course websites is in Appendix A.

4. Course Titles, Goals, and Objectives

Collectively the following different course titles are used in the courses we were able to find on the web (see Appendix A for complete information). The numbers in the parenthesis, if any, are the repeat count.

- Information Retrieval (7)
- Information Retrieval and Data Mining
- Information Retrieval and Search Engines
- Information Retrieval and Web Agents
- Information Retrieval and Web Search (3)
- Intelligent Information Retrieval
- Intelligent Information Retrieval and Web Search
- Introduction to Information Retrieval
- Modern Information Retrieval
- Search Engines and Web Navigation

- Search Engine Development
- Search Engine Technologies
- Web Based Information Architecture
- Web Data Mining
- Web Information Retrieval and Management
- Web Information Search
- Web Search and Data Mining
- Web Search Engines
- WWW Search Engines Algorithms, Architectures and Implementations

As can be seen, if we remove peripheral phrases such as “introduction”, “intelligent”, and “modern”, and put phrases such as “web search” and “web search engine” into the same category, we could classify the course titles into three categories:

1. Courses that emphasize the theme of information retrieval;
2. Courses that emphasize the theme of web search and data mining; and
3. Courses that combine the above two themes.

If we consider “web search” and “data mining” as two very different categories, the courses probably could be divided into four groups, those concentrating on information retrieval, those on web search, those on data mining, and those that combine all three (information retrieval, web search, and data mining).

While the exact course goals and objectives vary from different courses, a subset of course goals and objectives are common. These common goals or objectives include theory, practice, and implementation of information retrieval and its application in web search. The common goals reflect the nature and the contents of the courses. Here are some quotes from the course websites regarding the goals or objectives of the courses.

- Look at the methods used to search for and retrieve information from collections of documents, including Web search systems and library catalogs. The course combines theoretical and practical approaches, and includes sections on user interfaces and evaluation of the effectiveness of information retrieval systems.[3]
- Study the theory, design, and implementation of text-based information systems.[8]
- Discuss the design of a Web search engine and the extraction of information off the Web.[10]
- Focus on the technologies for storing and retrieving large-scale hypertext datasets. Particular emphasis is given to the data structures and algorithms needed to build efficient search engines for the World Wide Web (WWW).[11]
- Introduce students to the principles of information storage and retrieval systems and databases.[13]
- Understand the how the Web is organized, understand the characteristics and limitations of web search, develop several of the components to implement a web search engine, and make a significant contribution to an open source search engine project.[17]
- Study the basic principles and practical algorithms used for those Web Information Systems.[27]
- Discuss theory and practice of searching and retrieval of text and bibliographic information.[28]

- Survey principles and techniques in information retrieval with a focus on text databases, including automatic indexing, search techniques, query mechanisms, relevance feedback, and evaluation methodology. Students will examine the performance of selected commercial and web-based systems.[31]

5. Topics Covered in the Courses

Since a course in the area of information retrieval and web search typically is an elective one, there are no required core components to cover, as one might find in other courses where the core is designated by the ACM and IEEE curriculum guidelines.[4] The exact topics vary from course to course, depending on the audience, the interests and expertise of the instructor(s), and other factors. Here we summarize the course topics in two groups, one focuses on the area of information retrieval, and the second focuses on search engines and related web technologies.

1. *Main topics of information retrieval:* Typical topics include text indexing, common retrieval models such as Boolean, vector, and probabilistic models, retrieval evaluation, query languages and operations, user modeling, and interface issues.
2. *Main topics of web search engines and technologies:* Typical topics include web search, crawling and indexes, link analysis, web meta-data, search engine architectures, web usage mining, spam and advertising, and social networks.

Most of the courses discuss a combination of the topics in the two main areas, information retrieval theory and its applications in software systems such as web search engines. A few courses are notably more tailored towards general information retrieval such as the one at CMU[8] and the one at UMass[2], while a few others are more explicitly on the subjects of web search such as the one at NYU[10] and the one at Harding University.[17]

6. Textbooks

Throughout the courses we surveyed, the textbook by Manning, Raghavan, and Schütze (MRS)[16] is by far the most popular one, 17 of the 28 courses use it as one of the main textbooks, including three universities in China. The other two popular books are the one by Baeza-Yates and Ribeiro-Neto (BYRN)[5] (eight of 28) and the one by Croft, Metzler, and Strohman (CMS)[9] (five of 28). Both MRS and BYRN concentrate on the topics in information retrieval in general. MRS presents a more recent treatment of the topics than those of the BYRN as it is dated in 1999. Though the authors of BYRN have a new version of their book in 2011, the courses in our survey all quoted the book in its 1999 version at the time of our survey.

The authors of MRS aim the book at introductory level of graduate and upper level undergraduate students. The book contains a total of 21 chapters, each of which, according to the authors, can be covered in about one lecture unit of 75 to 90 minutes. The first eight chapters cover the core of information retrieval which includes retrieval models, index construction, term weights, ranking computation, and evaluation of retrieval. The second part of the book deals with more advanced topics using the foundation built in the first eight chapters. Various topics are discussed in this part, such as query processing, language models, classification and clustering, matrix decomposition, link analysis and other web search engine basics. BYRN, like MRS, starts with chapters that cover the basics of IR. The topics discussed in BYRN that are not in MRS include parallel and distributed IR, user interface and visualization, multimedia IR, and digital

libraries. The book by CMS puts more emphasis on the science and engineering behind the application of the information retrieval in web search engines. The book uses web search engines as a vehicle to discuss the topics in IR. The book studies more algorithms and data structures related to information retrieval that are used in search engines. In addition to these three popular books, about 25 other books are used as main text or main reference in the surveyed the courses. See Appendix B for a complete list.

7. Projects and Exercises

We consider here projects to be programming exercises that can result in some usable software components realizing a functional goal. Based on this loose definition, we found that the projects in the courses surveyed, when the descriptions are available, can be roughly divided into three categories, the ones that build a complete search system (simple or complex), the ones that modify a part or parts of an existing search system, and the ones that create a piece of software that functions as a stand-alone program to process, rank, or do other work on a body of text, but otherwise not as a complete search system.

Projects that fall into the first category, building a complete search system using a high level programming language include Davis[10] in which students build a question-answer system using web content; Mihalcea[22] in which students build a search engine within the UNT domain; and Yarowsky[32] in which students can choose to build systems to find friends, to classify news articles, or to help shopping. Projects that fall into the second category, modifying or creating a component to work with an existing search system include Agichtein[1] in which students implement a ranking function for the Lucene open-source search engine; McCown[10] in which students revise parts of the existing search engine Nutch; Strzalkowski[18] in which students extend a selected component in Lucene such as term weight, page scoring, query expansion, and relevance feedback. Projects that fall into the third category include Callan[4] in which students can choose to build personalized PageRank component, or a text classifier using Naïve Bayes method; Allan[2] in which students are asked to write software to process, classify, index, and search a dataset from Enron employees' emails, the size of which is about half a million; Wilson[31] in which students build an indexer through a series of smaller programming exercises.

8. Conclusions

In this paper, we present the result of our survey and analysis of courses on the subject of information retrieval and web search offered by colleges and universities across the web, mainly in the United States and in China. We collected the information available on the web from about 40 courses from different colleges and universities, six of which are from Chinese universities and research institutes. We concentrate on the goals of the course, main contents covered, textbooks used, and the student projects. The course contents can be roughly divided into two segments, the core of information retrieval, and some advanced topics and the applications of information retrieval to systems such as search engines. Among the two dozens of different textbooks used as a main reference in the courses, three are standing out as the most popular ones, that is, MRS[16], CMT[9], and BYRN[5]. While MRS and BYRN both present a more broad perspective on the subject of information retrieval, CMT uses a search engine as the main thread

of application of information retrieval, thus, the book contains more computer science specific information such as data structures and algorithms. Student programming projects in general fall into one of three categories, those that develop a search engine from scratch, those that revise a part or parts of an existing search engine, and those that develop stand-alone software to accomplish a particular functionality of a information retrieval system. We believe the information presented in the paper will help design, implementation, or revision of courses on the subject of information retrieval with various target audience groups in mind, especially for students at the undergraduate level.

References

[1] Agichtein, E. *CS572: Information Retrieval and Web Search* (http://www.mathcs.emory.edu/~eugene/cs572/) at Emory University, Spring 2010. Accessed December 30, 2011.

[2] Allan, J. *Information Retrieval* (http://cs646.cs.umass.edu/) at University of Massachusetts, Fall 2010. Accessed December 30, 2011.

[3] Arms, W.Y. *Information Retrieval* (http://www.infosci.cornell.edu/courses/info430/2007fa/index.html) at Cornell University, Fall 2007. Accessed December 30, 2011.

[4] Association of Computing Machinery. *Computer Science Curriculum 2008* (http://www.acm.org//education/curricula/ComputerScience2008.pdf) . Accessed December 30, 2011.

[5] Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval.* Addison Wesley.

[6] BCS. (2007). *Proceedings of the First International Workshop on Teaching and Learning of Information Retrieval (TLIR 2007).* Available at: http://ewic.bcs.org/category/16371. Accessed January 4, 2012.

[7] BCS. (2008). *Proceedings of the Second International Workshop on Teaching and Learning of Information Retrieval (TLIR 2008).* Available at: http://ewic.bcs.org/category/16295. Accessed January 4, 2012.

[8] Callan, J. and Yang, Y. *11-741: Information Retrieval* (http://boston.lti.cs.cmu.edu/classes/11-741/index.html) at Carnegie Melon University, Spring 2011. Accessed December 30, 2011.

[9] Croft, W.B., Metzler, D. and Strohman, T. (2009). *Search Engines: Information Retrieval in Practice.* Addison Wesley.

[10] Davis, E. *G22.2580: Web Search Engines* (http://cs.nyu.edu/courses/fall07/G22.2580-001/index.html) at New York University, Fall 2007. Accessed December 30, 2011.

[11] Davison, B. *CSE345/445: WWW Search Engines Algorithms, Architectures and Implementations* (http://www.cse.lehigh.edu/~brian/course/2009/searchengines/) at Lehigh University, Spring 2009. Accessed December 30, 2011.

[12] Fernandez-Luna, J.M., Huete, J.F., MacFarlane, A., and Efthimiadis, E.N. (2009). Teaching and learning in information retrieval. *Information Retrieval.* 12:201–226.

[13] Giles, C.L. *IST 441: Information Retrieval and Search Engines* (http://clgiles.ist.psu.edu/IST441/index.html) at Penn State, Spring 2012. Accessed December 30, 2011.

[14] Levene, M. *Search Engines and Web Navigation* (http://www.dcs.bbk.ac.uk/~mark/webtech.html) at Birkbeck University of London, Fall 2011. Accessed December 30, 2011.

[15] Li, W.J. *Web Search and Mining* (http://www.cs.sjtu.edu.cn/~liwujun/course/wsm.html) at Shanghai Jiao Tong University, Fall 2011. Accessed December 30, 2011.

[16] Manning, C., Raghavan, P., and Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

[17] McCown, F. *COMP 475: Search Engine Development* (http://www.harding.edu/fmccown/classes/comp475-s09/) at Harding University, Spring 2009. Accessed December 30, 2011.

[18] McCown, F. (2010). Teaching web information retrieval to undergraduates. In *Proceedings of the 41st ACM Technical Symposium on Computer Science Education* (SIGCSE 2010), Mar 2010, Milwaukee, WI, pp. 87-91. doi:10.1145/1734263.1734294

[19] McNamee, P. *605.744: Information Retrieval* (http://www.apl.jhu.edu/~paulmac/ir.html) at Johns Hopkins University, Spring 2011. Accessed December 30, 2011.

[20] Meng, X. (2003). Putting information retrieval theory into practice—A web search engine project for an undergraduate computer science elective course. In *Proceedings of the 2003 American Society for Engineering Education Annual Conference & Exposition.*
[21] Meng, X. and Xing, S. (2011). Teaching Web Information Retrieval and Network Communication Technology to Non-Major Undergraduate Students. In *Proceedings of the 2011 ASEE Annual Conference and Exposi*tion, Vancouver, B.C., Canada.
[22] Mihalcea, R. *CSCE 5200 Information Retrieval and Web Search* (http://www.cse.unt.edu/~rada/CSCE5200/) at University of North Texas, Spring 2011. Accessed December 30, 2011.
[23] Mooney, R. *Intelligent Information Retrieval and Web Search* (http://www.cs.utexas.edu/users/mooney/ir-course/) at University of Texas at Austin, Fall 2011. Accessed December 30, 2011.
[24] Nayak, P. and Raghavan, P. *Information Retrieval and Web Search* (http://www.stanford.edu/class/cs276/) at Stanford, Spring 2011. Accessed December 30, 2011.
[25] Nie, J.Y. *Web Based Information Architectures* (http://net.pku.edu.cn/~wbia/2011/index.html) at Peking University, Spring 2011. Accessed December 30, 2011.
[26] Schutze, H. *Introduction to Information Retrieval* (http://www.ims.uni-stuttgart.de/lehre/teaching/2009-SS/ir/) at Stuttgart University, Summer 2009. Accessed December 30, 2011.
[27] Si, L. *CS 490 WIR: Web Information Retrieval and Management* (http://www.cs.purdue.edu/homes/lsi/CS490W_Fall_2011/CS490W.html) at Purdue University, Fall 2011. Accessed December 30, 2011.
[28] Strzalkowski, T. *CSI 550: Information Retrieval* (http://aquarius.ils.albany.edu/~minoo/csi550/) at University of Albany, Fall 2011. Accessed December 30, 2011.
[29] Ward, N. *CS 5319/4390: Search Engine Technologies* (http://www.cs.utep.edu/nigel/search/) at University of Texas at El Paso, Spring 2009. Accessed December 30, 2011.
[30] Weikum, G. *Information Retrieval and Data Mining* (http://www.mpi-inf.mpg.de/departments/d5/teaching/ws07_08/irdm/) at Max Planck Institute, Winter 2008. Accessed December 30, 2011.
[31] Wilson, G.V. *Information Retrieval* (http://www9.georgetown.edu/faculty/wilsong/IR/IR.html) at Georgetown University, Spring 2008. Accessed December 30, 2011.
[32] Yarowsky, D. *Information Retrieval and Web Agents* (http://www.cs.jhu.edu/~yarowsky/cs466.html) at Johns Hopkins University, Spring 2011. Accessed December 30, 2011.
[33] Zhu, L., & Tang, C. (2006). A module-based integration of Information Retrieval into undergraduate curricula. *Journal of Computing Sciences in Colleges*, 22(2), 288–294.

Appendix A: List of course websites in the courses surveyed

1. Agichtein, E. *CS572: Information Retrieval and Web Search* (http://www.mathcs.emory.edu/~eugene/cs572/) at Emory University, Spring 2010. Accessed December 30, 2011.
2. Allan, J. *Information Retrieval* (http://cs646.cs.umass.edu/) at University of Massachusetts, Fall 2010. Accessed December 30, 2011.
3. Arms, W.Y. *Information Retrieval* (http://www.infosci.cornell.edu/courses/info430/2007fa/index.html) at Cornell University, Fall 2007. Accessed December 30, 2011.
4. Callan, J. and Yang, Y. *11-741: Information Retrieval* (http://boston.lti.cs.cmu.edu/classes/11-741/index.html) at Carnegie Melon University, Spring 2011. Accessed December 30, 2011.
5. Davis, E. *G22.2580: Web Search Engines* (http://cs.nyu.edu/courses/fall07/G22.2580-001/index.html) at New York University, Fall 2007. Accessed December 30, 2011.
6. Davison, B. *CSE345/445: WWW Search Engines Algorithms, Architectures and Implementations* (http://www.cse.lehigh.edu/~brian/course/2009/searchengines/) at Lehigh University, Spring 2009. Accessed December 30, 2011.
7. Giles, C.L. *IST 441: Information Retrieval and Search Engines* (http://clgiles.ist.psu.edu/IST441/index.html) at Penn State, Spring 2012. Accessed December 30, 2011.
8. Levene, M. *Search Engines and Web Navigation* (http://www.dcs.bbk.ac.uk/~mark/webtech.html) at Birkbeck University of London, Fall 2011. Accessed December 30, 2011.

9. Li, W.J. *Web Search and Mining* (http://www.cs.sjtu.edu.cn/~liwujun/course/wsm.html) at Shanghai Jiao Tong University, Fall 2011. Accessed December 30, 2011.
10. McCown, F. *COMP 475: Search Engine Development* (http://www.harding.edu/fmccown/classes/comp475-s09/) at Harding University, Spring 2009. Accessed December 30, 2011.
11. McNamee, P. *605.744: Information Retrieval* (http://www.apl.jhu.edu/~paulmac/ir.html) at Johns Hopkins University, Spring 2011. Accessed December 30, 2011.
12. Mihalcea, R. *CSCE 5200 Information Retrieval and Web Search* (http://www.cse.unt.edu/~rada/CSCE5200/) at University of North Texas, Spring 2011. Accessed December 30, 2011.
13. Mooney, R. *Intelligent Information Retrieval and Web Search* (http://www.cs.utexas.edu/users/mooney/ir-course/) at University of Texas at Austin, Fall 2011. Accessed December 30, 2011.
14. Nayak, P. and Raghavan, P. *Information Retrieval and Web Search* (http://www.stanford.edu/class/cs276/) at Stanford, Spring 2011. Accessed December 30, 2011.
15. Nie, J.Y. *Web Based Information Architectures* (http://net.pku.edu.cn/~wbia/2011/index.html) at Peking University, Spring 2011. Accessed December 30, 2011.
16. Schutze, H. *Introduction to Information Retrieval* (http://www.ims.uni-stuttgart.de/lehre/teaching/2009-SS/ir/) at Stuttgart University, Summer 2009. Accessed December 30, 2011.
17. Si, L. *CS 490 WIR: Web Information Retrieval and Management* (http://www.cs.purdue.edu/homes/lsi/CS490W_Fall_2011/CS490W.html) at Purdue University, Fall 2011. Accessed December 30, 2011.
18. Strzalkowski, T. *CSI 550: Information Retrieval* (http://aquarius.ils.albany.edu/~minoo/csi550/) at University of Albany, Fall 2011. Accessed December 30, 2011.
19. Ward, N. *CS 5319/4390: Search Engine Technologies* (http://www.cs.utep.edu/nigel/search/) at University of Texas at El Paso, Spring 2009. Accessed December 30, 2011.
20. Weikum, G. *Information Retrieval and Data Mining* (http://www.mpi-inf.mpg.de/departments/d5/teaching/ws07_08/irdm/) at Max Planck Institute, Winter 2008. Accessed December 30, 2011.
21. Wilson, G.V. *Information Retrieval* (http://www9.georgetown.edu/faculty/wilsong/IR/IR.html) at Georgetown University, Spring 2008. Accessed December 30, 2011.
22. Yarowsky, D. *Information Retrieval and Web Agents* (http://www.cs.jhu.edu/~yarowsky/cs466.html) at Johns Hopkins University, Spring 2011. Accessed December 30, 2011.

The following sites are from Chinese universities and research institutes, some of which are in English, some in Chinese, and some in mixed English and Chinese.

23. Du, X. *Intelligent Information Retrieval* (http://iir.ruc.edu.cn/courses/iir2010.jsp) at Renmin University. Spring 2010. (site in mixed English and Chinese, lecture notes in Chinese). Accessed December 30, 2011.
24. Li, R. and Lu, Z. *Modern Information Retrieval* (http://idc.hust.edu.cn/~rxli/teaching/ir.htm) at College of Computer Science and Technology, Huazhong University of Science and Technology (date unknown, site in English). Accessed December 30, 2011.
25. Li, W.J. Web *Search and Mining* (http://www.cs.sjtu.edu.cn/~liwujun/course/wsm.html) at Shanghai Jiao Tong University. Fall 2011. (site in English). Accessed December 30, 2011.
26. Nie, J.Y. *Web Based Information Architectures* (http://net.pku.edu.cn/~wbia/2011/index.html) at Peking University. Spring 2011. (site in English). Accessed December 30, 2011.
27. Wan, X. *Web Data Mining* (http://www.icst.pku.edu.cn/lcwm/course/WebDataMining/?page_id=2) at Peking University. Fall 2011. (site in Chinese). Accessed December 30, 2011.
28. Wang, B. *Modern Information Retrieval* (http://ir.ict.ac.cn/ircourse/) at Institute of Computing Technology of Chinese Academy of Sciences. Fall 2011. (site in Chinese). Accessed December 30, 2011.

Appendix B: A complete list of textbooks referenced in the courses surveyed

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval.* Addison Wesley.

Baldi, P., Frasconi, P., and Smyth, P. (2003). *Modelling the Internet and the Web*. John Wiley and Sons.

Battelle, J. (2005). *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*. Portfolio Hardcover.

Belew, R. K. (2001). *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW*. Cambridge University Press.

Bourne, C. P. and Hahn, T.B. (2003). *A History of Online Information Services: 1963-1976.* The MIT Press.

Buettcher, S., Clarke, C.L.A., and Cormack, G.V. (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press.

Chakrabarti, S. (2002). *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann.

Chang, G. (2001). *Mining the World Wide Web*. Springer.

Cheong, F. (1996). *Internet Agents: Spiders, Wanderers, Brokers, and Bots.* Indianapolis, IN : New Riders.

Croft, W.B., Metzler, D. and Strohman, T. (2009). *Search Engines: Information Retrieval in Practice*. Addison Wesley.

Croft, W. B. (ed). (2000). *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval*. Kluwer Academic Publishers.

Frakes, W. and Baeza-Yates, R. (1992). *Information Retrieval: Data Structures and Algorithms*. Englewood Cliffs, N.J. : Prentice Hall.

Greengrass, E. (2000). *Information Retrieval: A Survey.* Available online at: http://www.csee.umbc.edu/csee/research/cadip/readings/IR.report.120600.book.pdf.

Grossman, D. (n.d.). *Information Retrieval*. Available online at: http://ir.iit.edu/~dagr/cs529/ir_book.html

Han, J. and Kamber, M. (2000). *Data Mining - Concepts and Techniques*. Morgan Kaufmann.

Hearst, M. (2009). *Search User Interfaces*. Cambridge University Press. Available online at: http://searchuserinterfaces.com/

Hersh, W. R. (2003). *Information Retrieval: A Health and Biomedical Perspective*. 2nd Edition. Springer-Verlag.

Korfhage, R. R. (1997). *Information Storage and Retrieval*. John Wiley & Sons.

Kowalski, G. and Maybury, M. T. (2000). *Information Storage and Retrieval Systems: Theory and Implementation*. Kluwer Academic Publishers.

Langville, A. N. and Meyer, C.D (2006). *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton Press.

Levene, M. (2005). *An Introduction to Search Engines and Web Navigation*. Pearson.

Liu, B. (2011). *Web Data Mining.* Springer.

Manning, C., Raghavan, P., and Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Marchionini, G.(1997). *Information Seeking in Electronic Environments*. Cambridge University Press.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Available online at: http://www.dcs.gla.ac.uk/Keith/Preface.html

Salton, G. (1988). *Automatic text processing: the transformation, analysis, and retrieval of information by computer.* Reading, Mass. :Addison-Wesley.

Sparck Jones, K. and Willett, P. (1997). *Readings in Information Retrieval.* Morgan Kaufmann.

van der Weide, T. (2001). *Information Discovery*. Available online book at: http://osiris.cs.kun.nl/iris/web-docs/edu/ir1/ir1.pdf

Witten, I. H., Moffat, A., and Bell, T.C. (1999). *Managing Gigabytes.* Available online at: http://ww2.cs.mu.oz.au/mg/

Wong, C. (1997). *Web Client Programming.* O'Reilly and Associates. Available online at: *http://oreilly.com/openbook/webclient/*

Appendix C: Course goals, objectives, and topics which are accessible from the web

Professor Eugene Agichtein. *CS572: Information Retrieval and Web Search* at Emory (spring 2010)

- Basic and advanced techniques for text-based information systems: text indexing; Boolean, vector space, and probabilistic retrieval models; evaluation, feedback, user modeling, and interface issues; Web search including crawling, link-based algorithms, and Web metadata; text/Web clustering, classification; information extraction and text mining; Web2.0: Social networks, community systems, and user-generated content.

Professor James Allan *CMPSCI 646: Information Retrieval* at UMass (fall 2010)

- Core topics include material necessary to understand how an IR system is constructed and functions -- in particular, the material needed to carry out the class programming assignments. The following topics will be covered, though the order will be determined in part by student interest and class discussion: Evaluation, Retrieval models, Statistics of text, Indexing models, File organization, Efficiency, possibly including compression, Clustering, Relevance feedback, Document filtering, Distributed retrieval, Web search, Question answering, Multimedia retrieval, Cross-language retrieval, Advanced evaluation issues, Interactive retrieval, Interaction with Natural Language Processing

Professor William Y. Arms. *Information Retrieval* at Cornell (fall 2007)

- This course looks at the methods used to search for and retrieve information from collections of documents, including Web search systems and library catalogs. The course combines theoretical and practical approaches, and includes sections on user interfaces and evaluation of the effectiveness of information retrieval systems.
- Dictionaries, inverted files, postings, Term weighting, Similarity, ranking and the vector space model, String processing 1: Wild cards, stemming, and spelling, String processing: String search, Relevance feedback and query refinement, Latent semantic indexing, Probabilistic information retrieval, Evaluation of retrieval effectiveness, Web crawling, Architecture of information retrieval systems, Links and anchor text, Spam and advertising, Interfaces for browsing and searching, Metadata, Classification and categorization.

Professors Jamie Callan and Yiming Yang. *11-741: Information Retrieval* at CMU (spring 2011)

- This course studies the theory, design, and implementation of text-based information systems. The Information Retrieval core components of the course include statistical characteristics of text, representation of information needs and documents, several important retrieval models (Boolean, vector space, probabilistic, inference net, language modeling, link analysis), clustering algorithms, collaborative filtering, automatic text categorization, and experimental evaluation. The software architecture components include design and implementation of high-capacity text retrieval and text filtering systems.

Professor Ernest Davis. *G22.2580: Web Search Engines* at New York University (fall 2007)

- We will discuss the design of a Web search engine and the extraction of information off the Web. Topics include: Web crawlers, Database design, Query language, Relevance ranking, Document

Similarity and Clustering, The "invisible" Web, Specialized search engines, Evaluation, Natural Language Processing, The structure of the web, Intelligent retrieval and the semantic Web, Web content mining, Web usage mining, Multi-media retrieval, Multilingual retrieval.

Professor Brian Davison. *CSE345/445: WWW Search Engines Algorithms, Architectures and Implementations* at Lehigh University (Spring 2009)

- This course focuses on the technologies for storing and retrieving large-scale hypertext datasets. Particular emphasis is given to the data structures and algorithms needed to build efficient search engines for the World Wide Web (WWW). Topics covered include: information retrieval (IR) models, performance evaluation, query languages and operations, properties of hypertext, crawling, indexing, searching, ranking, link analysis, parallel and distributed IR, and user interfaces. Students will participate in class projects involving both the creation and management of a large document collection on the WWW. These projects will require programming in languages such as Perl/CGI, C/C++, or Java.

Professor C. Lee Giles. *IST 441: Information Retrieval and Search Engines* at Penn State (spring 2012)

- This course will introduce students to the principles of information storage and retrieval systems and databases. Students will learn how effective information search and retrieval is interrelated with the organization and description of information to be retrieved. Students will also learn to use a set of tools and procedures for organizing information, will become familiar with the techniques involved in conducting effective searches of print and online information resources and will build a vertical/specialty search engine.

Professor Mark Levene. *Search Engines and Web Navigation* at Birkbeck University of London (fall 2011)

- Course content: The structure of the web - **web metrics**, Find information on the web - **search and navigation**, How people use the web - **web data mining**.

Professor Frank McCown. *COMP 475: Search Engine Development* at Harding (spring 2009)

- The purpose of this class is understand the how the Web is organized, understand the characteristics and limitations of web search, develop several of the components to implement a web search engine, and make a significant contribution to an open source search engine project.

Professor Paul McNamee. *605.744: Information Retrieval* at Johns Hopkins (spring 2011)

- This course covers the storage and retrieval of unstructured digital information. Topics include automatic index construction, retrieval models, textual representations, efficiency issues, search engines, text classification, and multilingual retrieval.

Professor Rada Mihalcea. *CSCE 5200 Information Retrieval and Web Search* at University of North Texas (spring 2011)

- This course will cover traditional material, as well as recent advances in Information Retrieval (IR), the study of indexing, processing, and querying textual data. Basic retrieval models, algorithms, and IR system implementations will be covered. The course will also address more advanced topics in "intelligent" IR, including Natural Language Processing techniques, and "smart" Web agents.

Dr. Raymond Mooney. *Intelligent Information Retrieval and Web Search Course* at UT Austin (fall 2011)

- This course will cover traditional material as well as recent advances in information retrieval (IR), the study of the processing, indexing, querying, organization, and classification of textual documents, including hypertext documents available on the world-wide-web.

Professors Pandu Nayak and Prabhakar Raghavan *CS 276: Information Retrieval and Web Search* at Stanford (spring 2011)

- Basic and advanced techniques for text-based information systems: efficient text indexing; Boolean and vector space retrieval models; evaluation and interface issues; Web search including crawling, link-based algorithms, and Web metadata; text/Web clustering, classification; text mining.

Professor Hinrich Schutze. *Introduction to Information Retrieval* at Stuttgart (summer 2009)

- Boolean retrieval, The term vocabulary and postings lists, Dictionaries and tolerant retrieval, Index construction, Index compression, Scoring, term weighting and the vector space model, Computing scores in a complete search system, Evaluation in information retrieval, Text classification and Naive Bayes, Vector space classification, Flat clustering, Hierarchical clustering, Web search basics, Web crawling and indexes, Link analysis

Professor Luo Si. *CS 490 WIR: Web Information Retrieval and Management* at Purdue (fall 2011)

- This course studies the basic principles and practical algorithms used for those Web Information Systems. The contents include: Web search, recommendation system, Web information extraction, etc. The course emphasizes both the above applications and solid modeling techniques that can be extended for other applications. Students will: Learn the techniques behind Web search engines, E-commerce recommendation systems, etc.; Get hands on project experience by developing real-world applications, such as building a small-scale Web search engine, a Web page management system, or a movie recommendation system; Learn tools and techniques to do research in the area of information retrieval or text mining; Lead to the amazing job opportunities in Search Technology and E-commerce companies such as Google, Microsoft, Yahoo! and Amazon.

Dr. George V. Wilson. *Information Retrieval* at Georgetown (spring 2008)

- The course is a survey of principles and techniques in information retrieval with a focus on text databases, including automatic indexing, search techniques, query mechanisms (Boolean queries, topic hierarchies, natural language queries), relevance feedback, and evaluation methodology. Students will examine the performance of selected commercial and web-based systems.

Dr. David Yarowsky. *Information Retrieval and Web Agents* at Johns Hopkins University (spring 2011)

- **Information Retrieval** - Topics include a comprehensive study of current document retrieval models, mail/news routing and filtering, document clustering, automatic indexing, query expansion, relevance feedback, user modelling, information visualization and usage pattern analysis.
- **Text Understanding** - This segment of the course will focus on additional language processing steps for template filling and information extraction from retrieved documents, including reference resolution, sense tagging and summarization. Emphasis will be placed on recent, primarily statistical methods.
- **Web Agents and WWW Applications** - The final segment of the course will explore current issues in information retrieval and data mining on the World Wide Web. It will focus on case studies of web agents, spiders, robots and search engines, exploring both their practical implementation and the economic and legal issues surrounding their use. One of the hot technologies of the 21st century!