UNIVERSIDADE DA CORUÑA

Teaching Guide				
Identifying Data				2014/15
Subject (*)	Recuperación da información e web semántica		Code	614502010
Study programme	Mestrado Universitario en Enxeñaría Informática (plan 2012)			
Descriptors				
Cycle	Period	Year	Type	Credits
Official Master's Degree	1st four-month period	First	Obligatoria	6
Language	Spanish			
Prerequisites				
Department	ComputaciónTecnoloxías da Información e as Comunicacións			
Coordinador	Barreiro Garcia, Álvaro	E-mail	alvaro.barreiro@udc.es	
Lecturers	Barreiro Garcia, Álvaro	E-mail	alvaro.barreiro@udc.es	
	Cacheda Seijo, Fidel		fidel.cacheda@udc.es	
	Parapar López, Javier		javier.parapar@udc.es	
	Vázquez Naya, José Manuel		jose.manuel.vazquez.naya@udc.es	
Web				
General description	Os modelos, técnicas e algoritmos de recuperación de información estudados nesta materia permitirán aos estudantes comprender a arquitectura dos Search Engines para a web. Ademais os contidos prácticos da mesma capacitaranlles para construír os seus propios buscadores para traballar sobre repositorios de documento ou a web. Ademais durante os últimos anos houbo un interese crecente en idear unha web semántica a partir de meta-datos e anotacións. Unha web baseada en documentos xml e tags, meta-datos e esquemas, sen dúbida facilitaría os enormes retos aos que se enfronta a recuperación de información web. Nesta materia abórdanse tamén os modelos, técnicas e algoritmos de maior impacto desenvolvidos nos últimos anos co obxectivo de materializar unha web semántica. A Recuperación de Información en grandes coleccións de documentos e na web expón enormes retos (volume de datos, datos distribuídos, alta porcentaxe de datos volátiles, datos non estruturados e redundantes, heteroxeneidade, calidade dos datos e confianza) e a Web Semántica parte xa do gran reto da extracción de información cando os meta-datos non son expostos publicamente e expón novos retos como os do matching de ontologías, resolución de entidades ou unha dificultade maior en canto á heteroxeneidade e calidade dos datos e á indexación e procura semántica. Por todo iso a Recuperación de Información e a Web semántica constitúen un dos campos de mellores saídas profesionais en informática con oportunidades de negocio e emprego non só nas grandes compañías de Search Engines senón tamén en moitas pequenas e medianas compañías.			

Study programme competences	
Code	Study programme competences
A5	Capacidade de comprender e saber aplicar o funcionamento e organización da internet, as tecnoloxías e protocolos de redes de nova xeración, os modelos de compoñentes, sóftware intermediario e servizos.
A12	Capacidade para aplicar métodos matemáticos, estatísticos e de intelixencia artificial para modelar, deseñar e desenvolver aplicacións, servizos, sistemas intelixentes e sistemas baseados no coñecemento.
B1	Capacidade de resolución de problemas.
B5	Habilidades de xestión da información.
B10	Capacidade para proxectar, calcular e deseñar produtos, procesos e instalacións en todos os ámbitos da enxeñaría informática
B13	Capacidade para o modelado matemático, cálculo e simulación en centros tecnolóxicos e de enxeñaría de empresa, particularmente en tarefas de investigación, desenvolvemento e innovación en todos os ámbitos relacionados coa Enxeñaría en Informática
B14	Capacidade para a elaboración, planificación estratéxica, dirección, coordinación e xestión técnica e económica de proxectos en todos os ámbitos da Enxeñaría en Informática seguindo criterios de calidade e ambientais
B17	Capacidade para a aplicación dos coñecementos adquiridos e de resolver problemas en contornas novas ou pouco coñecidos dentro de contextos máis amplos e multidisciplinares, sendo capaces de integrar estes coñecementos
B21	Posuír e comprender coñecementos que acheguen unha base ou oportunidade de ser orixinais no desenvolvemento e/ou aplicación de ideas, a miúdo nun contexto de investigación
B22	Que os estudantes saiban aplicar os coñecementos adquiridos e a súa capacidade de resolución de problemas en contornas novas ou pouco coñecidos dentro de contextos máis amplos (ou multidisciplinares) relacionados coa súa área de estudo

B23	Que os estudantes sexan capaces de integrar coñecementos e enfrontarse á complexidade de formular xuízos a partir dunha información que, sendo incompleta ou limitada, inclúa reflexións sobre as responsabilidades sociais e éticas vinculadas á aplicación dos seus coñecementos e xuízos
B25	Que os estudantes posúan as habilidades de aprendizaxe que lles permitan continuar estudando dun modo que haberá de ser en gran medida autodirixido ou autónomo
C4	Desenvolverse para o exercicio dunha cidadanía aberta, culta, crítica, comprometida, democrática e solidaria, capaz de analizar a realidade, diagnosticar problemas, formular e implantar solucións baseadas no coñecemento e orientadas ao ben común.
C6	Valorar criticamente o coñecemento, a tecnoloxía e a información dispoñible para resolver os problemas cos que deben enfrontarse.
C7	Asumir como profesional e cidadán a importancia da aprendizaxe ao longo da vida.
C8	Valorar a importancia que ten a investigación, a innovación e o desenvolvemento tecnolóxico no avance socioeconómico e cultural da sociedade

Learning outcomes			
Subject competencies (Learning outcomes)	Study programme competences		
Know, understand and analyze different models of information retrieval and semantic web, techniques for their efficient implementation and their evaluation methodology.	AJ5		CJ6 CJ8
Know, understand and analyze the software platforms used to create these systems.	AJ5		CJ6 CJ7 CJ8
Design and build new systems or improve the existing ones.	AJ5 AJ12	BJ1 BJ5 BJ10 BJ13 BJ14 BJ17 BC1 BC2 BC5	CJ6 CJ7
Plan and perform the evaluation of information retrieval and semantic web systems. Analyze the evaluation results of the systems in order to improve their efficiency and effectiveness.	AJ5	BJ1 BJ5	CJ6 CJ7
Be able to treat correctly the ethical, privacy, confidentiality and security aspects of these systems.		BC3	CJ4 CJ6

Contents	
Topic	Sub-topic
Introduction	Information Retrieval and Search Engine Architecture
Information gathering	Crawling and feeds.
Text and Web page processing	Text pre-processing and parsing. Anchor text and Web link analysis, internationalization.
Indexes and ranking.	Building and compressing indexes. Efficient query processing.
Query formulation and results presentation.	Formulation and query re-writing. Snippets. Results visualization.
Information Retrieval Models.	Boolean, Vector-space, probabilistic, language models.
Evaluation	Evaluation of Information Retrieval Systems. Evaluation campaigns. Efficiency and effectiveness metrics. Evaluation design: training, test, statistical significance. Crowd-sourced evaluation.
Text mining.	Document clustering and classification
Distributed and Social search.	Federated search and distributed search. Blogs, micro-blogs and social networks.
Recommender systems	Collaborative filtering. Models and algorithms for recommendation. Recommender systems.

Planning			
Methodologies / tests	Ordinary class hours	Student?s personal work hours	Total hours
Workbook	1	15	16
Laboratory practice	20	30	50
Problem solving	4	12	16
Mixed objective/subjective test	2	18	20
Guest lecture / keynote speech	16	32	48
Personalized attention	0		0
(*)The information in the planning table is for guidance only and does not take into account the heterogeneity of the students.			

Methodologies	
Methodologies	Description
Workbook	Readings in order to consolidate and complement the knowledge and skills acquired.
Laboratory practice	Labs assignments dealing with development platforms in commercial use (Lucene, Terrier, Nutch, Jena, Protege, Pellet)
Problem solving	Problems and short questions to consolidate the contents presented in the master classes.
Mixed objective/subjective test	Test about the fundamental contents of the subject.
Guest lecture / keynote speech	The student will attend to the lectures given by the teacher about the different techniques, models and algorithms related to Information Retrieval and the Wemantic Web. The teacher will employ different levels of abstraction-detail and will guide the student in the fundamental and complementary readings.

Personalized attention	
Methodologies	Description
Laboratory practice Problem solving	Control of the development of the labs assignment in the allocated lab hours, and the teacher will pay special attention to the student in particularly difficult problems, if necessary.

Assessment		
Methodologies	Description	Qualification
Laboratory practice	Control of the labs assignments and evaluation of the results achieved.	50
Mixed objective/subjective test	Questions related to the knowledge acquired. Questions that involve reasoning over the knowledge acquired, that involve practical problem-solving on real life issues related to Information Retrieval and the Semantic Web.	50

Assessment comments
Aqueles estudantes con matrícula a tempo parcial ou calquer circunstancia xustificada que impida a asistencia as clases, deberán contactar cos docentes para determinar alternativas ao seguimento e avaliación da materia.

Sources of information

Basic	- Bob DuCharme (2011). Learning SPARQL. O'Reilly - C.D. Manning, P. Raghavan, H. Schutze. (2008). Introduction to Information Retrieval. Cambridge University Press - R. Baeza-Yates and B. Ribeiro-Neto. (2011). Modern Information Retrieval (second edition) . Addison Wesley/Pearson Education - F. Cacheda, J.M. Fernández, J. Huete (eds.) (2011). Recuperación de Información. Un enfoque práctico y multidisciplinar. Ra-Ma - W.B. Croft, D. Metzler, T. Strohman. (2009). Search Engines. Information Retrieval in Practice. Pearson Education - John Hebeler, Matthew Fisher, Ryan Blace, Andrew Perez-Lopez, Mike Dean. (2009). Semantic Web Programming. Wiley
Complementary	

Recommendations
Subjects that it is recommended to have taken before
Subjects that are recommended to be taken simultaneously
Subjects that continue the syllabus
Other comments

(*)The teaching guide is the document in which the URV publishes the information about all its courses. It is a public document and cannot be modified. Only in exceptional cases can it be revised by the competent agent or duly revised so that it is in line with current legislation.