

Materia: Análisis y Recuperación de Información

Área: Ingeniería de Software

Carga Horaria: 60 hrs.

Docente: Dra. Daniela Lis Godoy

Programa de Contenidos:

Unidad I: Introducción

Introducción a análisis y recuperación de información (RI). Componentes y procesos de un sistema de RI. Sistemas de RI en la Web. Modelos básicos de recuperación: modelo booleano, modelo probabilístico y modelo de espacio de vectores. Medidas de similitud y ranking de documentos. Extensiones de los modelos básicos.

Unidad II: Análisis y Recuperación de Información Textual

Representación de texto. Reducción de dimensionalidad. Selección de términos: técnicas estadísticas, lingüísticas y basadas en conocimiento. Stop-words. Algoritmos de stemming: eliminación de sufijos, variedad de sucesores, diccionarios y n -grams. Ley de Zipf. Identificación de Idioma. Uso de diccionarios semánticos. Funciones de asignación de pesos a términos. Procesamiento de consultas: optimización y expansión. Indexación automática: estructuras y algoritmos de búsqueda.

Unidad III: Análisis de Links

La Web como grafo. Relación entre texto e hyperlinks. Redes sociales y análisis de co-citaciones. Nociones de Hubs y Authorities. Algoritmos Page Rank, HITS y variaciones. Inferencia de comunidades Web a partir de la topología de links.

Unidad IV: Clasificación de Documentos

Definición y aplicaciones. Métodos de RI aplicados a clasificación: feedback de relevancia y algoritmo de Rocchio, clasificadores TF-IDF (term frequency/inverse document frequency). Aprendizaje inductivo de clasificadores: naïve Bayes, inducción de árboles de decisión, k -NN, razonamiento basado en casos (CBR – Case-Based Reasoning), CBR textual y otros algoritmos. Evaluación de la clasificación. Métricas: exactitud, precisión, tasa de error y otras medidas.

Unidad V: Clustering de Documentos

Definición y aplicaciones. Caracterización de algoritmos de clustering. Algoritmos basados en particionamiento. Algoritmos de clustering jerárquico aglomerativos y divisivos. Algoritmos de clustering conceptual. Evaluación del clustering. Métricas externas e internas: entropía, pureza, F-Measure, cohesividad y otras medidas.

Unidad VI: Agentes para Recuperación de Información

Introducción a agentes: definición y aplicaciones a RI. Enfoques de construcción. Enfoque basado en contenido: aprendizaje de perfiles de usuario, adaptación y predicción. Enfoques de filtrado colaborativo. Filtrado centrado en usuarios: comparación de perfiles, identificación de vecinos y predicción. Filtrado centrado en ítems: coeficientes de correlación, similitud y predicción. Filtrado basado en modelos: clustering de usuarios y/o ítems, reglas de asociación. Enfoques híbridos.

Unidad VII: Evaluación y Experimentación

Métricas de performance en recuperación de información: precision, recall, F-measure y otras medidas. Colecciones estándar de documentos para evaluación de sistemas de recuperación, filtrado, clasificación, clustering y filtrado colaborativo de información.

Bibliografía

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

Croft, W. B.: *What Do People Want from Information Retrieval?: The Top 10 Research Issues for Companies That Use and Sell IR Systems*. D-Lib Magazine, November 1995.

Jain, A. K., Murty, M. N. and Flynn, P. J.: *Data Clustering: A Review*. ACM Computing Surveys, 31(3): 264-323, 1999.

Maes, P.: *Agents that Reduce Work and Information Overload*. Communications of the ACM, 37(7): 30-40, 1994.

Klusch, M.: *Information Agent Technology for the Internet: A Survey*. Data & Knowledge Engineering, 36(3): 337-372, 2001.

Korfhage, R.: *Information Storage and Retrieval*. New York: John Wiley & Sons, 1997.

Liu, B.: *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*. Springer, 2007.

Sebastiani, F.: *Machine Learning in Automated Text Categorization*. ACM Computing Surveys, 34(1): 1-47, 2002.

Sparck Jones, K. and Willett, P.: *Readings in Information Retrieval*. San Francisco: Morgan Kaufmann, 1997.

Forma de Evaluación:

La forma de evaluación es a través de un trabajo en el cual los alumnos aplican las técnicas de análisis y recuperación de información vistas en la materia a un dominio y aplicación determinados.