

Web Mining e Retrieval (WMR)

Docente: Roberto Basili

Informatica: Crediti 9 Anno 2 Semestre 2 CdL Specialistica

Il Web è la più grande collezione di informazione in formato digitale attualmente disponibile in modo pubblicamente accessibile. Il corso affronta gli aspetti teorici e realizzativi che ne consentono lo sfruttamento, dai processi di indicizzazione, accesso e recupero di informazione alla acquisizione di conoscenza da grandi collezioni di dati distribuite geograficamente. Le finalità del corso sono di:

- Approfondire tematiche legate all'apprendimento automatico, presentando i metodi avanzati di induzione di conoscenza dai dati.
- Conoscere i diversi modelli utilizzati nei motori di ricerca per il WWW e nelle loro declinazioni semantiche (Semantic Enterprise Search).
- Conoscere le tecnologie avanzate di Intelligenza Artificiale applicata al Web, per il trattamento linguistico dei testi (Natural Language Processing) e sperimentarne la applicazione nei domini del Social Web in problemi di Semantic document management, Link Analysis e Opinion Mining.

Programma

Sezione I: **Machine Learning e Learning basato su kernel.**

Richiami. Metodi Supervised. Metodi probabilistici e generativi. Metodi *Unsupervised*. *Clustering*. Metriche di similarità semantica. Metodi agglomerativi. *K-mean*.

Modelli Markoviani. Hidden Markov Models.

Kernel-based Learning. Kernel polinomiali e RBF. String Kernels. Tree kernels. Latent Semantic kernels. Semantic kernels. Applicazioni.

Sezione II: **Statistical Language Processing**

Supervised Language Processing tools. HMM-based POS tagging. Named Entity Recognition. Statistical parsing. PCFGs: Charniak parser. Modelli di Parsing Lessicalizzati. Shallow Semantic Parsing: kernel based semantic role labeling. Information Extraction.

Sezione III: **Web Mining & Retrieval.**

Modelli di *ranking* per il Web. Introduzione alla *Social Network Analysis*: rango, centralità.

Modelli di random walk: Page Rank. Motori di ricerca. SEO. Google.

Sistemi di *Question Answering*. *Open-domain Information Extraction*.

Acquisizione di Conoscenza da Wikipedia. *Social Web*. Algoritmi su grafi per la *community detection*.

Introduzione all'*Opinion Mining* e al *Sentiment Analysis*.

Testi

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. [Consultabile on-line](#)
- C.M. Bishop "Pattern Recognition and Machine Learning" Springer, 2006
- Roberto Basili, Alessandro Moschitti, *Text Categorization: from Information Retrieval to Support Vector Learning*, ARACNE Editore, 2005.
- Bing Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. 2nd Edition, July 2011, Springer.
- Note del docente e articoli scientifici distribuite durante il corso.

Aula e ora lezioni

Informatica: TBD (2° semestre)

Ricevimento studenti

Dopo una delle lezioni settimanali previste. Inoltre un'ora a settimana presso lo studio del docente

Comunicazioni

Per una migliore fruizione del Corso è fortemente consigliata la sua associazione con i Corsi di *Machine Learning e Information Retrieval del Semestre 1*. E' inoltre fortemente consigliata la fruizione del Corso di *Natural Language Processing* durante lo stesso Semestre 2.

Strumenti e tool per la progettazione di sistemi di Web retrieval basati su Machine learning verranno resi disponibili durante il Corso in lezioni di laboratorio dedicate. In progetti dedicati verranno progettate e sperimentate piattaforme per task di *Statistical Natural Language Processing*, *Link Analysis* ed *Opinion Mining* nell'ambito dei *Web-based Content Management Systems*.