

Web Mining, Search Engines, and Business Intelligence

A First Course in Web Mining, Search Engines, and Business Intelligence

Telephony & Shared Communications

HDTV and Telecommunications



About Dr. Garcia

Dr. Garcia research interests include Web Mining, Search Engine Architectures, and Information Retrieval at the intersection of Information Security and Intelligence. He is a program committee member of W3C's Adversarial Information Retrieval on the Web Workshops (AIRWeb) and has served as reviewer for *JASIST*, *IBM's Computer and Graphics*, and has co-chaired several local conferences on search engine technologies. At Polytechnic University, he is a visiting lecturer. He is the founder of <http://www.miislita.com>, an online resource on information retrieval and search engine technologies.

Description: This is a hands-on, one-semester course on Web Mining, search engines, and business intelligence. Students will learn by doing: (a) how search engines index and rank web documents, (b) how to conduct business intelligence from online resources, and (c) how to apply Web Mining strategies and algorithms in their research or workplace.

Target: Students in Business, Engineering, and Computer Sciences and from other disciplines are encouraged to register for this special course.

Requirements: Calculus II or Permission from advisor or department.

Grading: Take-home work and a final exam.

Topics: The following topics will be covered, not necessarily in this order:

- **Document Indexing:** Indexing of Web sites and text operations used by search engines including document linearization, tokenization, stop word filtration, stemming, and parsing.
- **Search Engine Optimization:** Mining search engine relevance algorithms for ranking high Web pages.
- **Intelligence Searching:** Covers undocumented (smart) searches in Google and other search engines; includes Hacking and Penetration through customized searches.
- **Keyword Research and Clustering:** Discovery of word patterns and keywords for branding and marketing through Association, Scalar, and Metric Clusters.
- **Term Matching Algorithms:** Vector Space Models used by search engines. Scoring of local, global, and entropy term weights.
- **Concept Matching Algorithms:** Singular Value Decomposition (SVD) and Latent Semantic Indexing (LSI) models for clustering and ranking.
- **Link Analysis Models:** Google's PageRank, Hubs & Authorities, and other link-based models.
- **Spam Intelligence:** Tools and techniques for spamming search engines and web sites. Includes techniques based on scripts, cloaking, keyword spam techniques, link-bombs, email marketing, viral marketing, Web 2.0, and Web 3.0.
- **Introduction to Business Dashboards (BDs):** Overview of dashboard technology, including open source, and customized add-on components.
- **Special Topics: On-Topic Analysis, Co-Occurrence Theory, and Latent Graphs.** This is my own area of research.

Textbook: Web Mining evolves on a daily basis; thus, there is no official textbook. However, the following reference books are recommended for research. Additional references will be provided in class.

1. *Modern Information Retrieval* (Baeza-Yates and Ribeiro-Neto; Addison Wesley).
2. *Information Retrieval – Algorithms and Heuristics* (Grossman and Frieder; Springer).