

Web Based Information Architectures

复习要点

- 1) 掌握 WWW 的基本概念, 包括 web,html/sgml,URI,tcp/ip,http,DNS
- 2) 掌握 Crawling 的基本算法, General crawling 面临的主要问题, 并相应的解决技术办法。(scalable,fast,polite,robust,etc.)
- 3) 掌握 Focused Crawling 基本原理。包括两个主要方面 Topical Locality 和 Topic Relevance Function.
- 4) 掌握 Web 的 2 个基本属性: power law 和 small world。掌握一种 generative model, 它要解决的问题, 模型的思想。
- 5) 掌握信息检索 (Information Retrieval) 的三种基本模型 (boolean model, vector model, probabilistic model)。
- 6) 掌握信息检索系统的主要评估指标 (recall, precision, F value, precision at 11 standard recall levels)。
- 7) 了解 indexing 过程的目标、基本步骤。(tokenize, stopper, stemmer)。掌握 indexes 倒排文件组织方式, 使用倒排索引的检索算法, 比如 vsm 模型下的检索实际例题。
- 8) 掌握对 Web 进行链接分析的两种基本方法: Hits, Pagerank。
- 9) 掌握文本分类的一般过程与方法, 重点掌握 Rocchio 方法, kNN, 支持向量机 (SVM) 进行文本分类的思想方法。了解文本分类器的评估指标。
- 10) 掌握文本聚类的一般过程与方法, 重点掌握基于划分的 (k-means, k-medoids)、层次的(Agglomerative)聚类方法。
- 11) 潜在语义标引 (LSI/SVD) 是如何减少词频矩阵大小的? 列举一两个应用实例。