《**Web** 信息处理与应用》复习提纲

PART 1：Web Search

一． Introduction

1. Web 搜索的概念与挑战

2. 信息检索（IR）的概念、与Web 搜索之间的关系

3. IR 与DB 之间的区别

4. IR 的任务与基础性问题

二． Web Crawler

1. 网络爬虫的概念和基本过程

2. 网络爬虫的主要需求

3. 网络爬虫的常用策略

4. 网络爬虫涉及的协议

5. 分布式爬虫与一致性 Hash 的概念

三． Text Processing

1. 文本处理的概念

2. 分词、分词歧义、未登录词、停用词等概念

3. 中文分词的挑战

4. 常用的分词方法

5. 词根化（Stemming）和编辑距离的概念

四． Indexing

1． 布尔检索、关联矩阵的概念

2． 倒排索引：概念、结构、构建算法、存储等

五． Queries

1. 查询表达的难点

2. 相关性反馈：概念、基本过程

3. 相关性反馈的分类及其各自的概念与特点

4. Ricchio 算法

5. 查询扩展的概念

6. 查询扩展的几种方法

六． Ranking

1. Ranking 的难点

2. 信息检索模型的概念、分类

3. Jaccard 系数

4. tf、df、tf-idf 的概念与计算

5. 向量空间模型

6. 余弦相似度的定义

7. 概率模型的概念

8. PageRank

9. HITS

七． Evaluation

1. 信息检索评价概述

2. 信息检索评价指标的分类

3. Precision、Recall、F-measure 的定义

4. P@N、R@Precision、AP 的定义

5. MAP、MRR

6. NDCG

PART 2：Web Information Extraction

一． Named Entity Recognition

1. 信息抽取（IE）的概念以及与IR 的关系

2. MUC-7 定义的信息抽取任务

3. 信息抽取的内容

4. NER 的概念与难点

5. MUC-7 中定义的NER 内容

6. NER 的性能评价指标

7. NER 的常用方法

二． Relation Extraction

1. 关系抽取的概念和意义
2. 关系的表示方法
3. 关系抽取的常用方法

PART 3：Web Data Mining

一. 概述(Introduction)

1. 网络挖掘的概念，包含哪些方面的内容，分别有哪些重要应用？

二. 网络内容挖掘(Web Content Mining)

数据(Data)

1. 概念：数据对象(Objects)，属性(Attributes)，维度(Dimensions)，特征(features)
2. 高维诅咒(Curse of dimensionality)现象。
3. 对于数据的预处理有哪些方法？其中需要掌握采样(Sampling)，特征选择(Feature selection)及降维(Dimensionality reduction)的基本原理。

分类(Classification)

4. 监督学习(Supervised learning)与无监督学习(Unsupervised learning)的关系与区别。
5. 分类(Classification)的基本原理。
6. 数据的向量表示(Vector space representation)
7. 熟练掌握 k 近邻算法，包括影响算法性能的要素——近邻个数及距离（相似度）度量。
8. 熟练掌握 Logistic regression 分类方法。
9. 如何评价分类效果？理解训练错误率，测试错误率以及泛化错误率的区别。

聚类(Clustering)

10. 聚类(Clustering)的基本原理及准则。
11. 层次式聚类算法流程，两个类之间的距离定义。
12. 熟练掌握 K-means 算法——算法流程，优化目标，收敛性分析。
13. 聚类算法的评价标准。

三. 网络结构挖掘(Web Structure Mining)

1. 网络结构如何用图来表示？图的组成部分以及相关性质。

社区分析(Community)

2. 社区(Community)的概念

3. 社区发现与聚类的关系。

4. 如何计算结构相似度？

5. 图分析的一些重要矩阵：邻接(Affinity)矩阵，拉普拉斯(Laplacian)矩阵，以及它们的一些重要性质。

6. Cut 概念；ratio cut 以及 normalized cut 的定义及推导。

7. Modularity 概念及其推导。与 spectral clustering 的相同点及不同点。

影响力分析(Influence)

8. 几种度量节点中心性的标准。

9. 两种影响力传播模型——线性阈值模型(Linear Threshold Model)，层级传播模型(Independent Cascade Model)的传播过程及区别。

10. 最大影响节点集(Most influential set)——问题建模，贪心算法以及算法的近似度。

11. 子模性质(submodularity)。

四. 推荐系统

1. 推荐系统基本模型以及一般工作流程。

2. 基于内容的推荐算法流程及优缺点

3. 协同过滤推荐算法流程及优缺点