中国科学技术大学

《Web 信息处理与应用》复习提纲

PB10210016 徐波　2013.1.9

Instructor

Jin Pei-Quan（金培权）　　Xu Lin-Li　（徐林莉）

Email: jpq@ustc.edu.cn　　Email: linlixu@ustc.edu.cn

Teaching Assistants

林盛, Ph.D. student　　于永波, Master Student

Phone: 13485728758　　Phone: 13865979122

Email: linsh@mail.ustc.edu.cn　　Email: yyb2012@maiul.ustc.edu.cn

Room: 1610, 科技实验楼西楼　　Room: 1610, 科技实验楼西楼

《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. 字符编码：ASCII、Unicode、UTF－8
3. 分词、分词歧义、未登录词、停用词等概念
4. 中文分词的挑战
5. 常用的分词方法
6. 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 Mining

一． 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． 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. 熟练掌握最小二乘算法——推导过程，闭式解，规范化之后的求解推导。
9. 过拟合现象出现的原因。
10. 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)

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

三． 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． Web Recommendation

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

提纲总结

PART 1：Web Search

一． Introduction

1. Web 搜索的概念与挑战

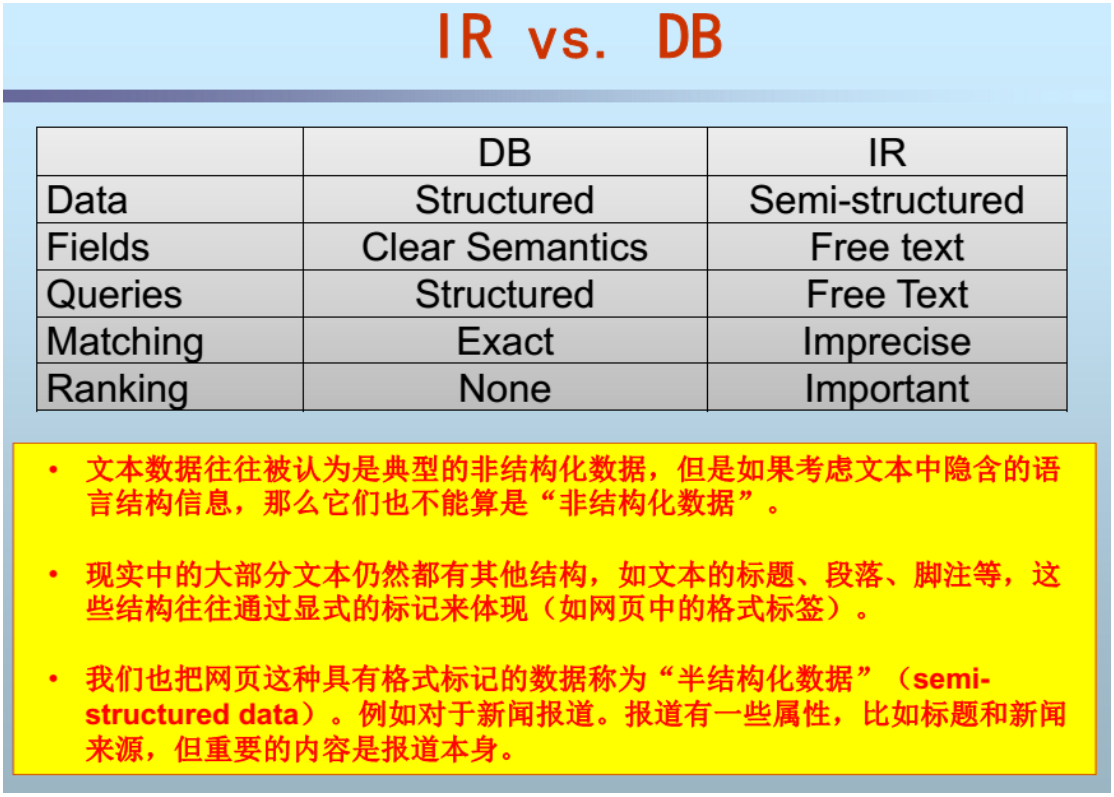
Web 搜索是指采用自动或半自动的方式，遵循一定的策略在 Web 上搜集和发现信息。实现 Web 搜索的技术统称为 Web 搜索技术，主要包括制定搜索策略、对网页超链接结构进行分析、评价 Web 信息资源的质量、分析信息资源的内容以及计算 Web 信息资源与搜索查询的相关程度等。

数据量增大、无结构信息、异构数据、数据的分布性、数据不稳定、数据的质量、异构数据、高价值信息的发现。

2. 信息检索（IR）的概念、与 Web 搜索之间的关系

信息检索（Information Retrieval）是指信息按一定的方式组织起来，并根据信息用户的需要找出有关的信息的过程和技术。狭义的信息检索就是信息检索过程的后半部分，即从信息集合中找出所需要的信息的过程，也就是我们常说的信息查寻（Information Search 或 Information Seek）。

3. IR 与 DB 之间的区别



4. IR 的任务与基础性问题

IR的任务

- 基于用户查询的搜索
  - 有时称为特殊搜索(ad hoc search)，因为查询的范围巨大而且事先没有约定）是搜索引擎研究的主要任务
- 过滤(filtering)
- 分类(classification)
- 问答(question answering)
  - “珠穆朗玛峰的高度是多少？”、“亚马逊河流有多长？”

表1-1 信息检索的维度

内容实例	应用实例	任务实例
文本	网络搜索	特殊搜索
图像	垂直搜索	过滤
视频	企业搜索	分类
扫描文档	桌面搜索	问答
音频	P2P搜索	
音乐		

基础性问题;
相关性计算、检索模型、评价、信息需求、检索功能。

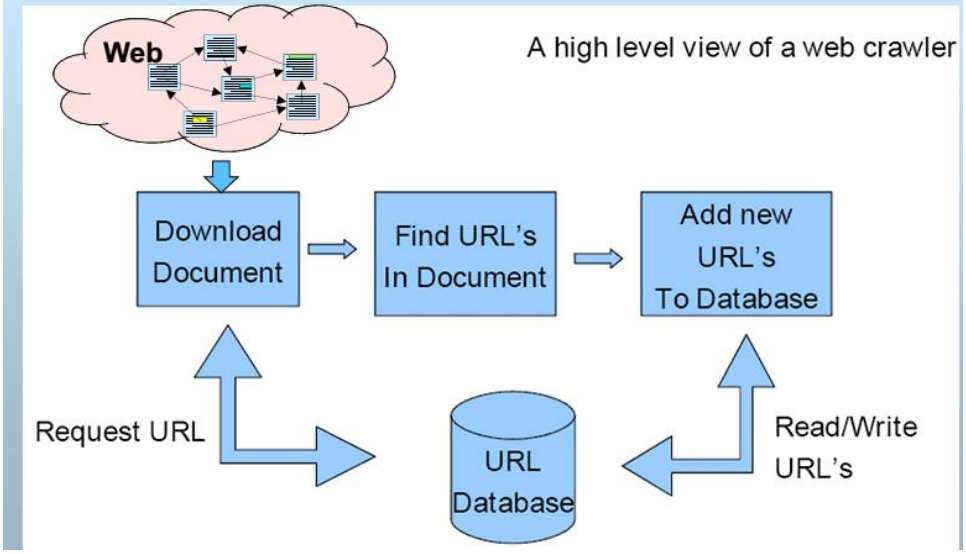
二． Web Crawler

1. 网络爬虫的概念和基本过程

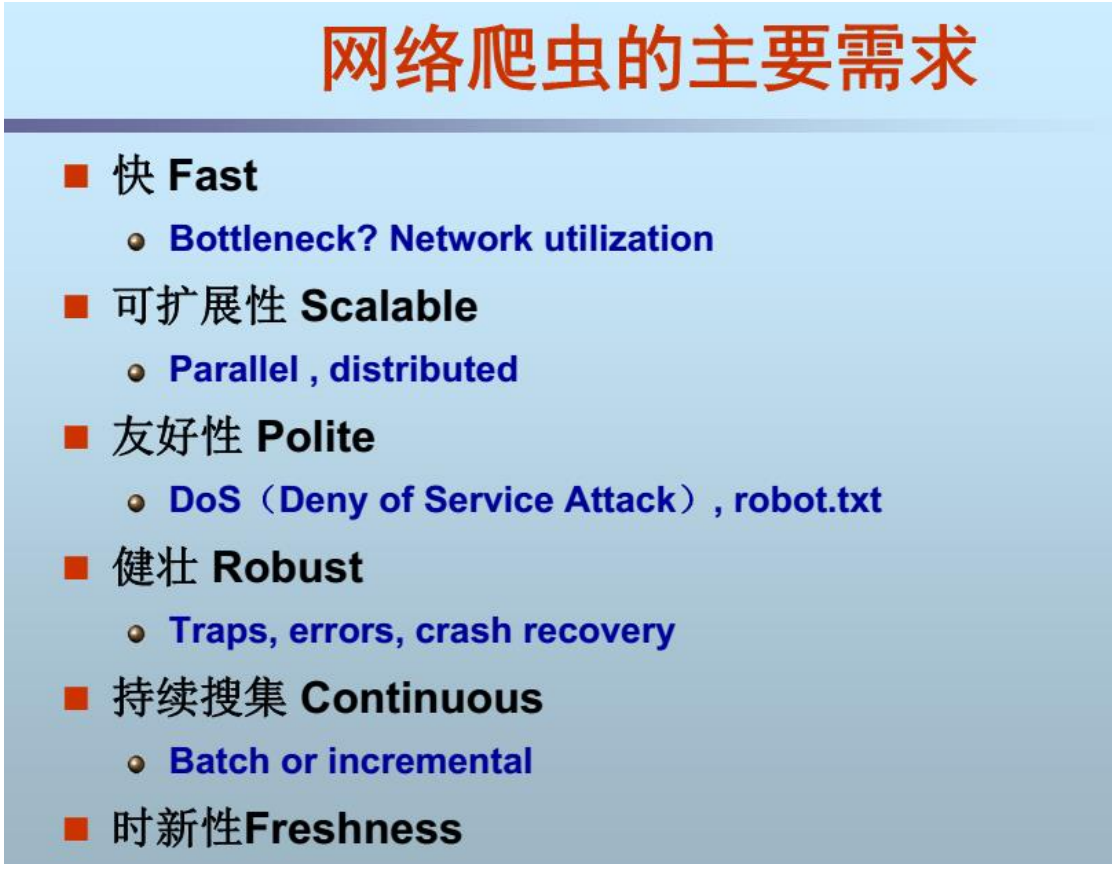
Web Crawler的任务定义

- 从一个种子站点集合（Seed sites）开始，从Web中寻找并且下载网页，获取排序需要的相关信息，并且剔除低质量的网页

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

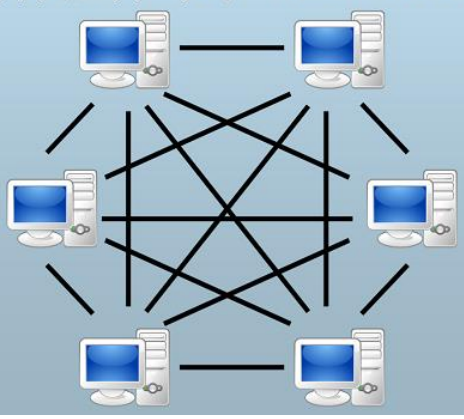
爬虫涉及的协议

- HTTP/HTML
- DNS/URL
- Robots Exclusion
- Sitemap

5. 分布式爬虫与一致性 Hash 的概念

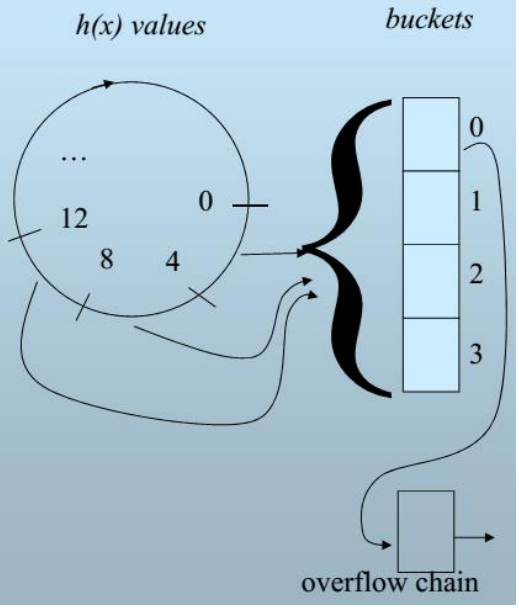
分布式爬虫

- M个节点同时执行搜集
- 问题：如何有效的把N个网站的搜集任务分配到M个机器上去？
- 目标：任务分配得均匀（ Balance ）



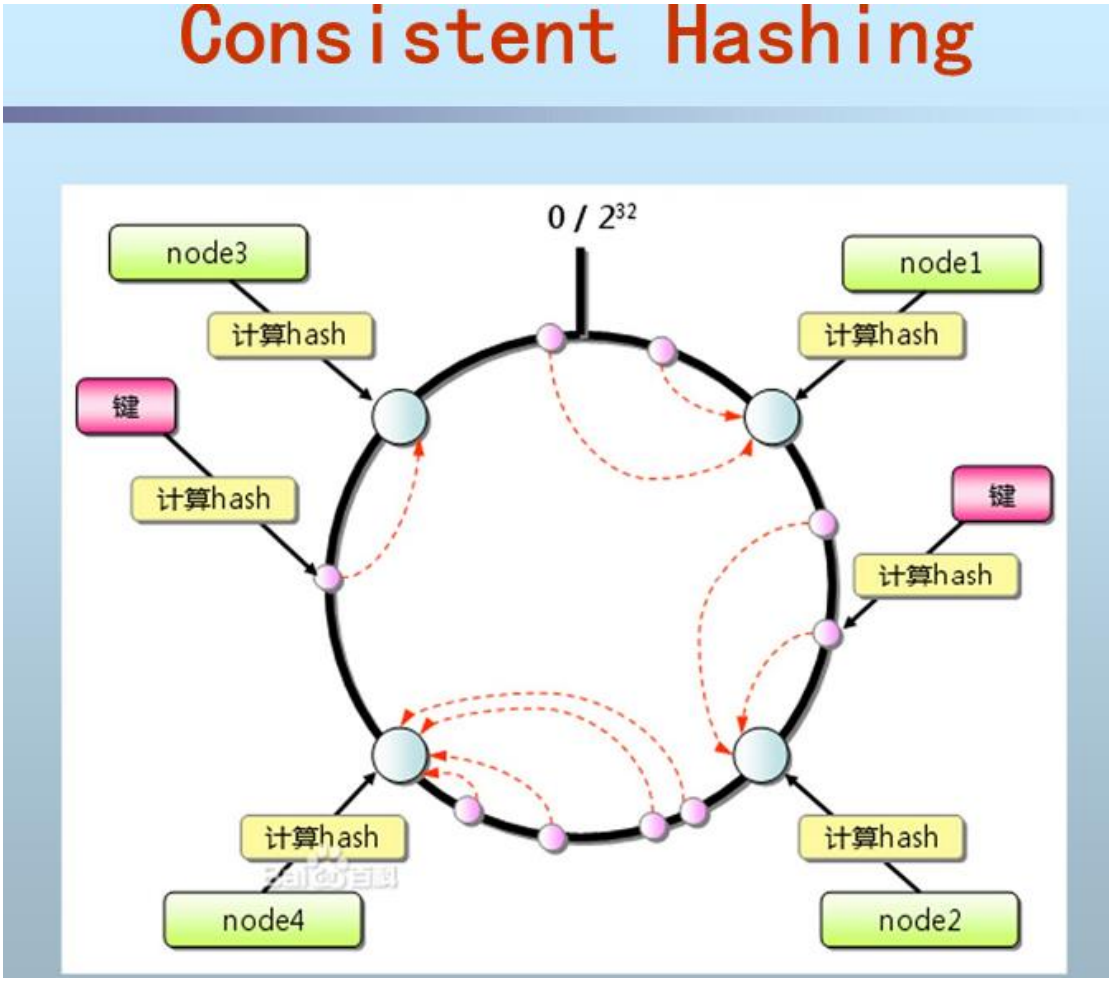
Hashing

- 从一个值均匀分布的 hash 函数开始:
  - h(name) → a value (e.g., 32-bit integer)
- 把 values 映射到 hash buckets
  - 一般取模 mod (# buckets)
- 把 items 放到buckets
  - 冲突，需要 overflow chain 解决冲突



Consistent Hashing

- 使用一个巨大的hash key空间
  - 2^{32} bits
  - 组织成回路
- URL和目标网站IP都hash映射到一个hash key空间，每个hash key对应一台抓取计算机
  - Node(url) or Successor(url)
- 如果某台抓取计算机失效，则该机器中的URL都迁移到顺时针方向的下一个node
  - 保证最小的数据迁移
  - 保证负载均衡，因为每个bucket都很小



三． Text Processing

1. 文本处理的概念

Text Processing

- Basic component in IR systems (not only for Web search).
- Also known as Document Processing
  - Converting raw documents into terms to be indexed
  - Enabling the matching of terms in the query to those in the documents.
- Document processing and query parsing are connected.

2. 字符编码：ASCII、Unicode、UTF－8

- **Character Encoding**
  - Bits <--> Characters 之间的映射方法，e.g.
    - ASCII: 100 0001 —— A
    - Unicode UTF-8: 11100111 10001000 10110001 ——爱
- **ASCII编码**
  - 1963提出，针对英文、数字、常用标点等
  - 1个字节（7 bits for characters，1 bit for error checking but usually not used）
  - 2^7 = 128 字符
- **ASCII对于许多语言来说远远不够**
  - 中文——5万多字，其中常用的有3000多
  - 泰米尔语——247种语言构造

Unicode （万国码、统一码）

- 1991 Unicode 1.0, 2012.9 Unicode 6.2
- 为所有语言提供统一的字符编码
  - 使电脑可以用统一、简单的方式来呈现和处理文字
- Unicode使用16进制的码位(codepoint)来表示字符
  - 码位：组成码空间的数值
  - ASCII的码位：0～7F
  - Unicode：包含1,114,112个码位, 0～10FFFF (Unicode 6.1)。通常用“U+4位16进制”表示，如：U+7231——爱
- 常见的Unicode编码方式
  - UTF-32: 每个码位固定使用4 bytes
  - UTF-8：每个码位使用1～4之间可变长度的 bytes

UTF-8 (8-bit Unicode Transformation Format)

- 用1到4个字节编码Unicode字符
- 若码位小于等于127，使用1 bytes
  - 与ASCII兼容，高位bit为0
- 若码位大于127，使用2～4 bytes
  - 第1个字节：由换码序列开始（连续的“1”并以“0”结束），例如“110”、“1110”。连续的“1”个数表示该码位使用的字节数。换码序列不计入字符的有效bits
  - 其余字节：由“10”开始，表示非ASCII字符，并且不计算入字符表示的有效bits

UTF-8的表示范围

10进制	Unicode 16进制	bit数	UTF-8	byte数
0－127	0000 0000 ~ 0000 007F	0~7	**0**XXX XXXX	1
128－2047	0000 0080 ~ 0000 07FF	8~11	**110**X XXXX **10**XX XXXX	2
2048－65535	0000 0800 ~ 0000 FFFF	12~16	**1110** XXXX **10**XX XXXX **10**XX XXXX	3
65536－1114111	0001 0000 ~ 001F FFFF	17~21	**1111 0**XXX **10**XX XXXX **10**XX XXXX **10**XX XXXX	4

UTF-8的表示示例

Unicode 16进制	Unicode 2进制	bit数	UTF-8 2进制
00CA	1100 1010	8	**110**0 0011 **10**00 1010
F03F	1111 0000 0011 1111	16	**1110**1111 **10**00 0000 **10**11 1111

3. 分词、分词歧义、未登录词、停用词等概念

- **Segmentation / Tokenization**
  - 将文档的字符串序列变成词序列
- 英文词语空格区分
  - **“University of Science and Technology of China”**
- 汉语、日语等无空格区分，分词困难
  - “中国科学技术大学”
  - “ニューヨーク大学”

相当于：UniversityofScienceandTechnologyofChina

交集型歧义（交叉歧义）

- 如果AB和BC都是词典中的词，那么如果待切分字串中包含“ABC”这个子串，就必然会造成两种可能的切分：“AB/C/”和“A/BC/”。这种类型的歧义就是交集型歧义。
- 比如“网球场”就可能造成交集型歧义
  - 网球/场/
  - 网/球场/

组合型歧义（组合歧义）

- 如果AB和A、B都是词典中的词，那么如果待切分字串中包含“AB”这个子串，就必然会造成两种可能的切分：“AB/”和“A/ B/ ”。这种类型的歧义就是组合型歧义。
- 比如“个人”就可能造成组合型歧义
  - 我/个人
  - 三/个/人/

未登录词即未包括在分词词表中但必须切分出来的词，包括各类专名（人名、地名、企业字号、商标号等）和某些术语、缩略词、新词等等

- “于大海发明爱尔肤护肤液”需要切分成“于大海/发明/爱尔肤/护肤液”，并需要识别出“于大海”是人名，“爱尔肤”是商标名，“护肤液”是术语名词。
- 如“斯普林菲尔德是伊里诺州首府”、“丹增嘉措70多岁了”，其中的美国地名、藏族人名都需识别。
- 机构名和商品品牌名：“希望电脑”、“国际乒联”、“非常可乐”。
- 专业领域的大量术语：“线性回归”、“A*算法”。
- 新词语，缩略语：“粉丝”、“E时代”、“坑爹”。

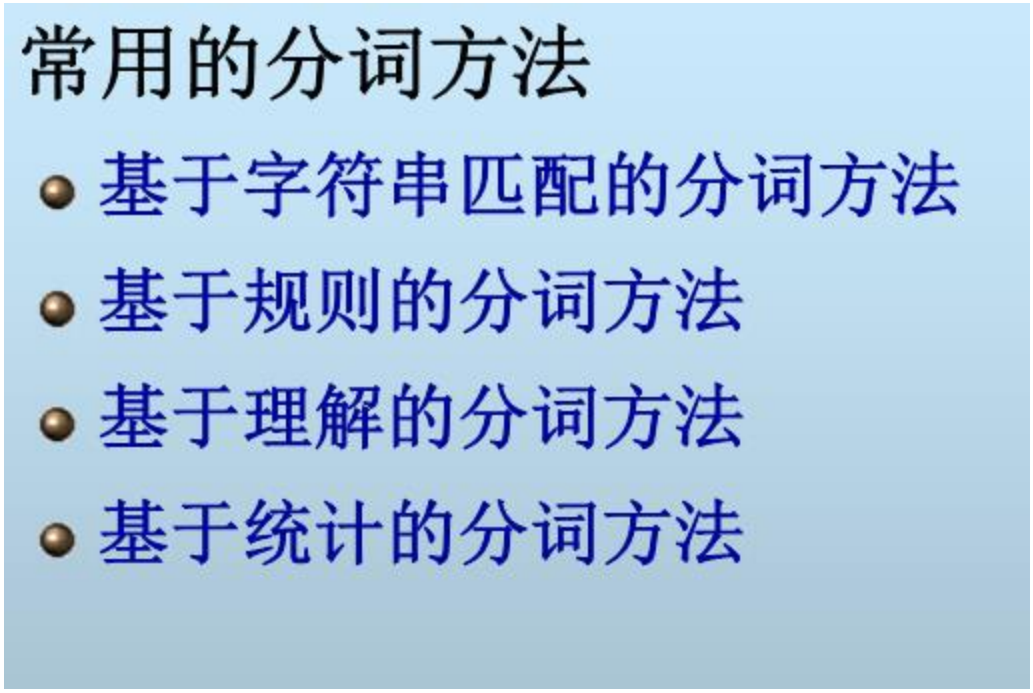
停用词－stopwords

- 在文档中频繁出现的词语
- 与语料库特性有关
  - 例如，wikipedia语料库中“wiki”是停用词

4. 中文分词的挑战

- 中文分词的挑战
  - 英语——词的集合 vs. 汉语——字的集合
  - 汉字之间存在着不同的组词方式
    - 如“发展中国家兔的饲养”一句，现有的汉语词就可能导致有两组语词分隔结果：发展中国家/兔/的/饲养，发展/中国/家兔/的/饲养。
  - 汉语虚词众多，而且绝大多数汉字当与不同的汉字组词时，其词可能为关键词，也可能为非用词
    - 如，“非”与“常”、“洲”分别组成不同意义的词“非常”（关键词）、“非常”（停用词）。
  - 分词歧义
  - 新词的频繁出现也给汉语分词增添了难度
    - 未登录词

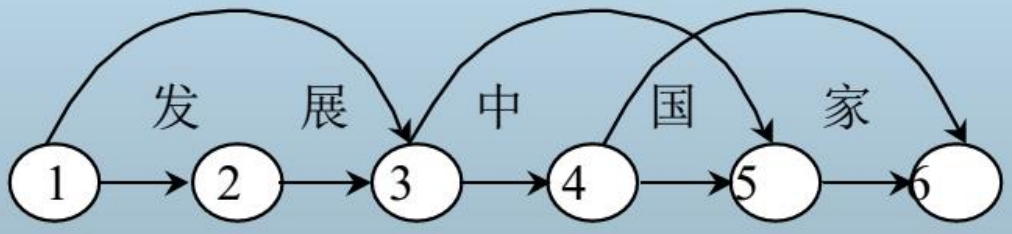
5. 常用的分词方法


常用的分词方法
基于字符串匹配的分词方法
基于规则的分词方法
基于理解的分词方法
基于统计的分词方法


1、基于字符串匹配的分词方法
常用的机械分词方法
正向最大匹配分词（FMM）
反向最大匹配分词（BMM，RMM）
双向最大匹配分词（BM：FMM＋RMM）

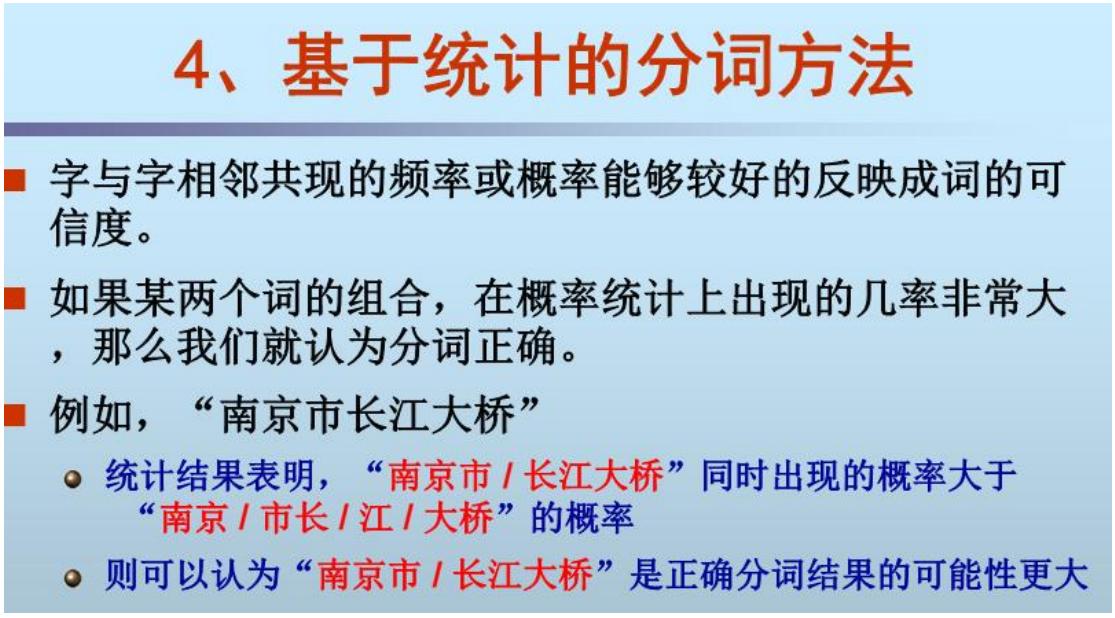
2、基于规则的分词方法

- 最少切分分词方法
  - 使句子中切出的词数目最少
  - 等价于在有向图中搜索最短路径问题



3、基于理解的分词方法

- 这种分词方法是通过让计算机模拟人对句子的理解，达到识别词的效果；
- 基本思想
  - 在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象。
- 需要使用大量的语言知识和信息。由于汉语语言知识的复杂性，难以将各种语言信息组织成机器可直接读取的形式，因此目前基于理解的分词系统还处在试验阶段



6. 词根化（Stemming）和编辑距离的概念

编辑距离

- Given two strings, *s* and *t*, the edit distance, or Levenshtein Distance, between them is the minimum number of ***edit operations*** required to transform *s* into *t*
- 例如
  - **Edit-Distance(“kitten”,“sitting”)=3**
    - **k**itten → **s**itten (substitution of "s" for "k")
    - sitt**e**n → sitt**i**n (substitution of "i" for "e")
    - sittin → sitting (insertion of "g" at the end).

四． Indexing

1． 布尔检索、关联矩阵的概念

最简单的检索方式：布尔检索

- 指利用 **AND, OR 或者 NOT**操作符将词项 连接起来的查询
  - 信息 **AND** 检索
  - 信息 **OR** *检索*
  - 信息 **AND** 检索 **AND NOT** 教材
- 在30多年中是最主要的检索工具
- 当前许多搜索系统仍然使用布尔检索模型
  - 电子邮件、文献编目、**Mac OS X Spotlight**工具

Term-Document 关联矩阵 (Incidence Matrix)

Terms \ Docs	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

1 if play contains word, 0 otherwise

2. 倒排索引：概念、结构、构建算法、存储等

二、倒排索引

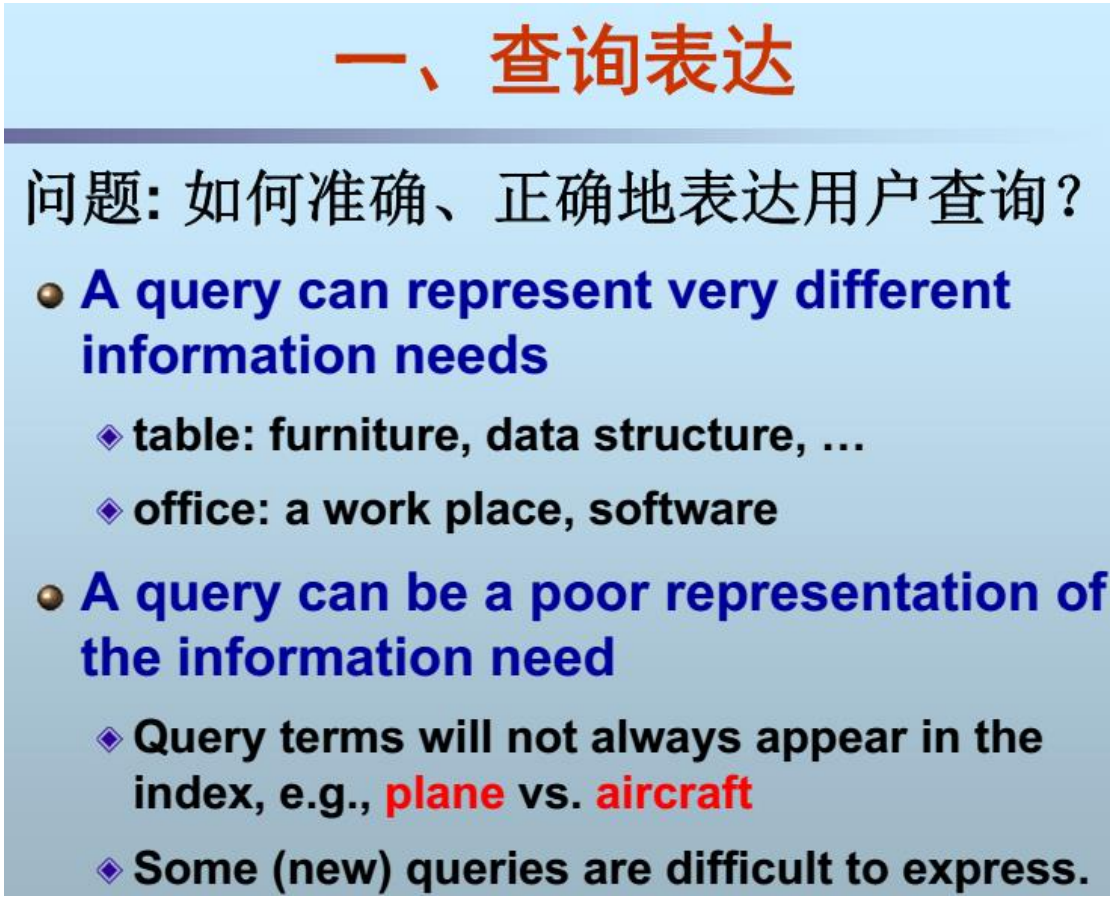
- IR中流行的基于词项的文本索引
- 包含两部分结构
  - Vocabulary (Dictionary) : a set of terms
  - Postings List: doc list where the term appeared

Vocabulary		Postings List
term1	→	Document17, Document 45123
		.
		.
termN	→	Document991, Document123001

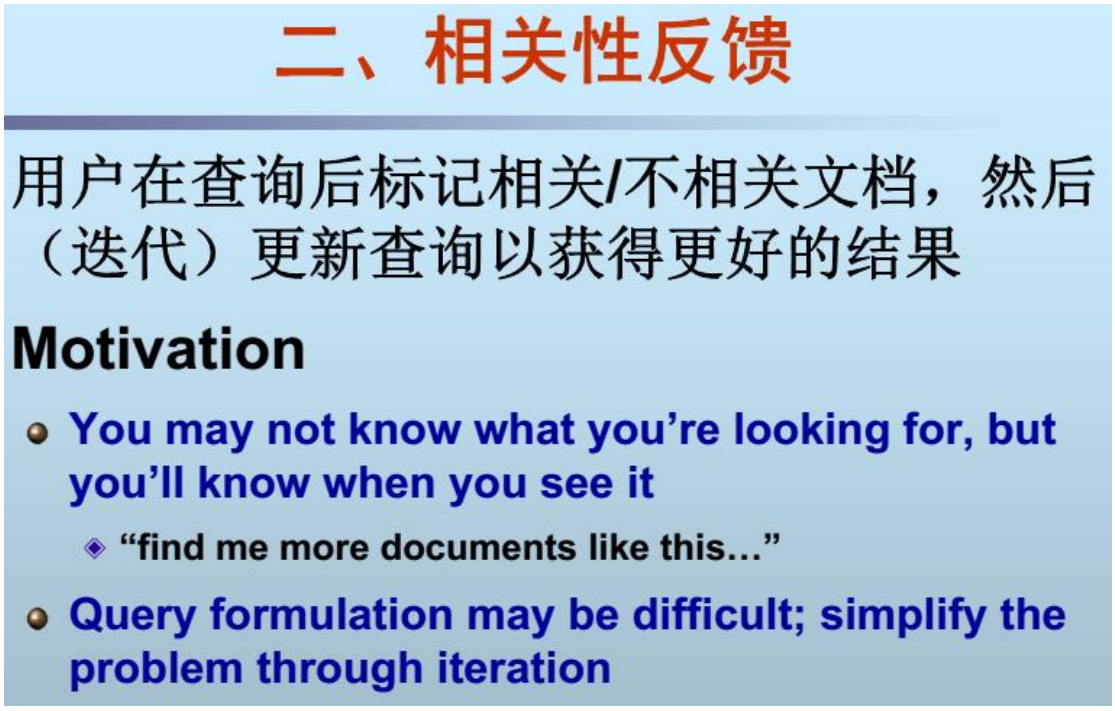
	doc. freq	trem. freq
new	2	1 [freq. =1]; 4 [freq. =1]
home	4	1 [freq. =1]; 2 [freq. =1]; 3 [freq. =1]; 4 [freq. =1];
sales	4	1 [freq. =1]; 2 [freq. =1]; 3 [freq. =1]; 4 [freq. =1];
top	1	1 [freq. =1];
forecasts	1	1 [freq. =1];
rise	2	2 [freq. =1]; 4 [freq. =1];
in	3	2 [freq. =1]; 3 [freq. =2];
july	3	2 [freq. =1]; 3 [freq. =1]; 4 [freq. =1];
increase	1	3 [freq. =1];

五． Queries

1. 查询表达的难点



2. 相关性反馈：概念、基本过程



二、相关性反馈

User issues a (short, simple) query

The user marks returned documents as relevant or non-relevant.

The system computes a better representation of the information need based on feedback.

Relevance feedback can go through one or more iterations.

Idea: it may be difficult to formulate a good query when you don't know the collection well, so iterate

通常用术语“ad hoc retrieval”来表示那种无相关反馈的常规检索

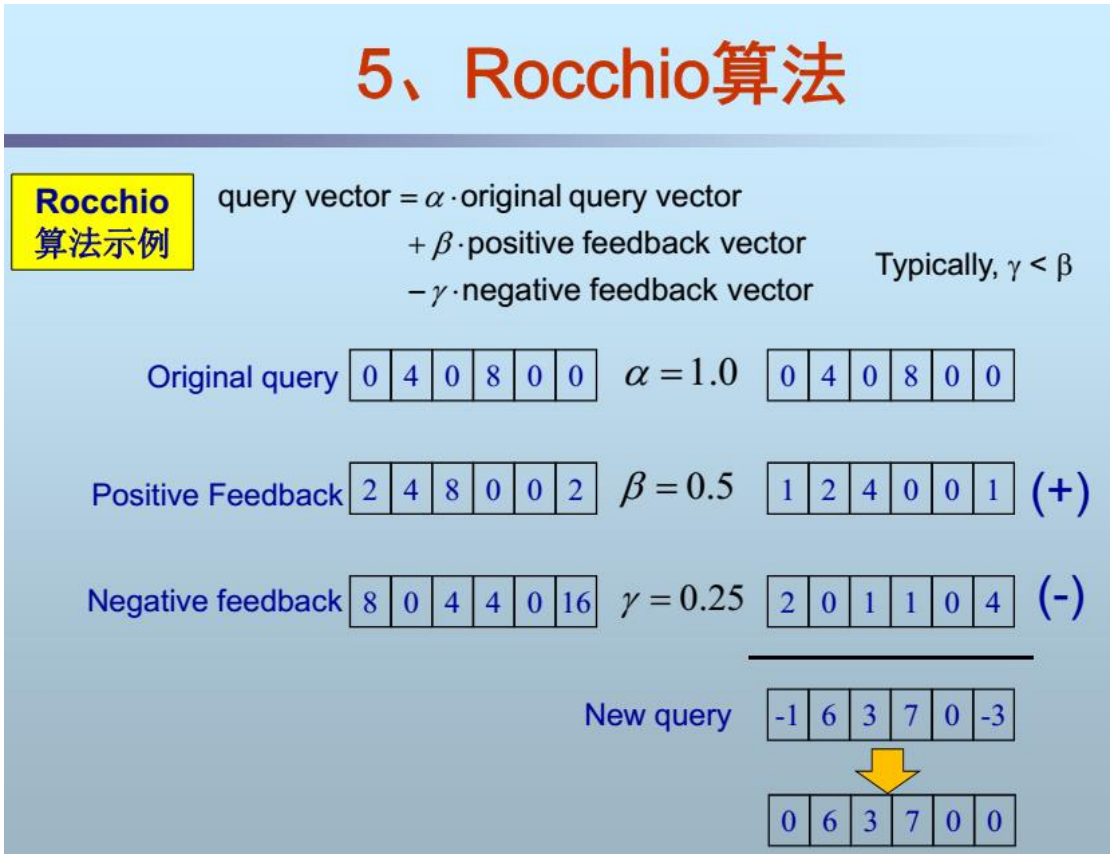
二、相关性反馈

相关性反馈如何工作？

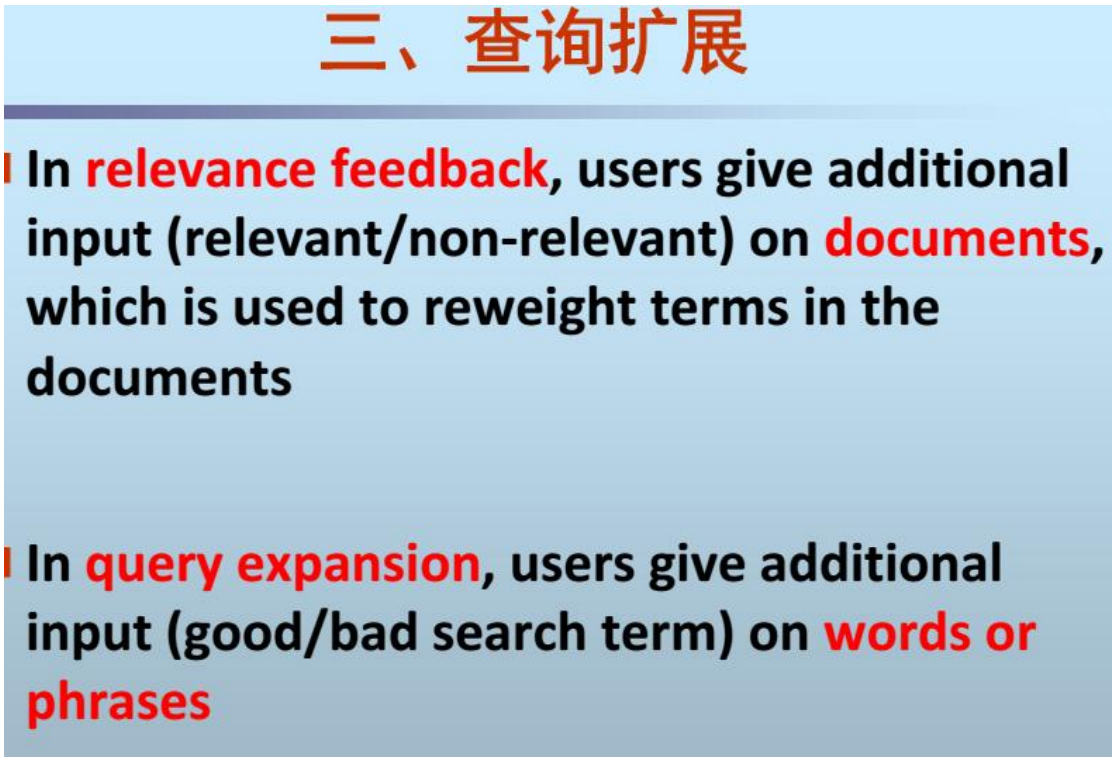
- Let's assume that there is an optimal query
  - The goal of relevance feedback is to bring the user query closer to the optimal query
- How does relevance feedback actually work?
  - Use relevance information to update query
  - Use query to retrieve new set of documents
- What exactly do we "feed back"?
  - Boost weights of terms from relevant documents
  - Add terms from relevant documents to the query
- Note that this is hidden from the user

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



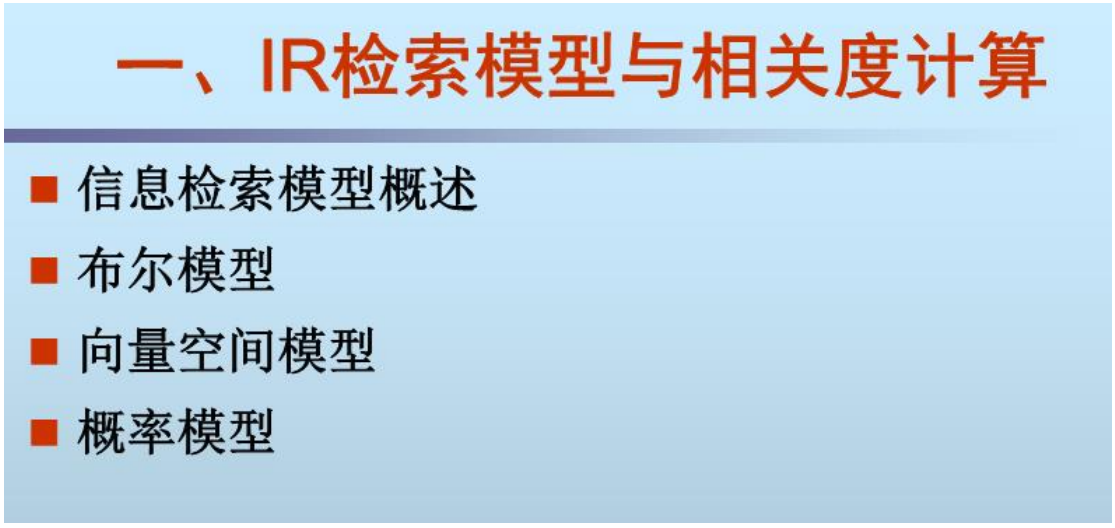
6. 查询扩展的几种方法

人工同义词典
Thesaurus 自动构建
基于搜索日志的查询扩展

六． Ranking

1. Ranking 的难点

2. 信息检索模型的概念、分类



3. Jaccard 系数

第一种方法：Jaccard系数

- 1901年Jaccard提出的计算两个集合重合度的常用方法
  - 令 A 和 B 为两个集合
  - Jaccard系数的计算方法:

$$\text{JACCARD}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

  - JACCARD (A, A) = 1
  - JACCARD (A, B) = 0 如果 A ∩ B = 0
- Jaccard 系数会给出一个0到1之间的值

Paul Jaccard (1868-1944)

查询 “ides of March”

文档 “Caesar died in March”

⇨ JACCARD(q, d) = 1/6

4. tf、df、tf - idf 的概念与计算

第四种方法：tf-idf

- 词项的tf-idf权重是tf权重和idf权重的乘积

$$w_{t,d} = (1 + \log \text{tf}_{t,d}) \cdot \log \frac{N}{\text{df}_t}$$

  - 随着词项频率的增大而增大
  - 随着词项罕见度的增加而增大
- 信息检索中最出名的权重计算方法
  - 注意：上面的 “-”是连接符，不是减号
  - 其他叫法：tf.idf、tf *idf、tfidf等

- 某个词项在A文档中出现100次，即tf = 100，在B文档中 tf = 10，那么A比B更相关
- 但是相关度不会相差10倍
- 相关度不会正比于词项频率tf

- df_t 是出现词项t的文档数目

5. 向量空间模型

3、向量空间模型

- 向量空间模型（Vector Space Model，VSM）是由G·Salton等人在1958年提出的。代表系统SMART
  - D={D1, D2,…}，Di=(Wi1 , Wi2 , …, Win)，Wij是词项的tf-idf权值
  - q= (Wq1 , Wq2 , …,Wqn)，Wqi是查询词项的tf-idf值
  - F：非完全匹配方式
  - R
    - 用文档和查询两个向量相似度来估计文档和查询的相关性
    - 文档和查询之间的相关度具有较强的可计算性和可操作性，不再只有0和1两个值

6. 余弦相似度的定义

3、向量空间模型

余弦相似度计算

$$\cos(\vec{q},\vec{d}) = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

- q_i 是第i 个词项在查询q中的tf-idf权重
- d_i是第i个词项在文档d中的tf-idf权重

7. 概率模型的概念

查询为：“Web 信息 教程”

- 所有词项在相关、不相关情况下的概率p_i、q_i分别为：

词项 / 概率	Web	信息	教材	教程	课件
R=1时的概率p_i	0.8	0.9	0.3	0.32	0.15
R=0时的概率q_i	0.3	0.1	0.35	0.33	0.10

文档D1： 信息 课件

则： P(D|R=1)=(1-0.8)*0.9*(1-0.3)*(1-0.32)*0.15

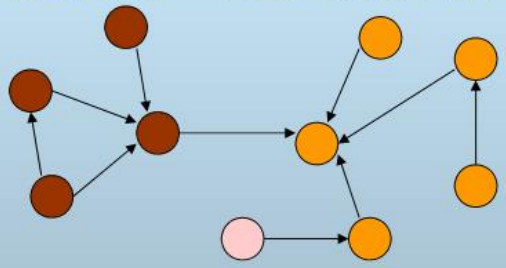
P(D|R=0)= (1-0.3)*0.1*(1-0.35)*(1-0.33)*0.10

P(D|R=1)/P(D|R=0)=4.216

8. PageRank

二、PageRank

- **PageRank算法：**
  - 将网页或者文档视作一个点，网页之间的超链接视作有向边，则会构成一个巨大的有向图。



  - 相同颜色代表主题相关网页（主题相关的点的连接要多于普通网页之间的连接），点之间的有向连接反映了网页之间互相引用，参考和推荐的关系，入度越多，则被引用或推荐的次数越多，网页的重要性就越大

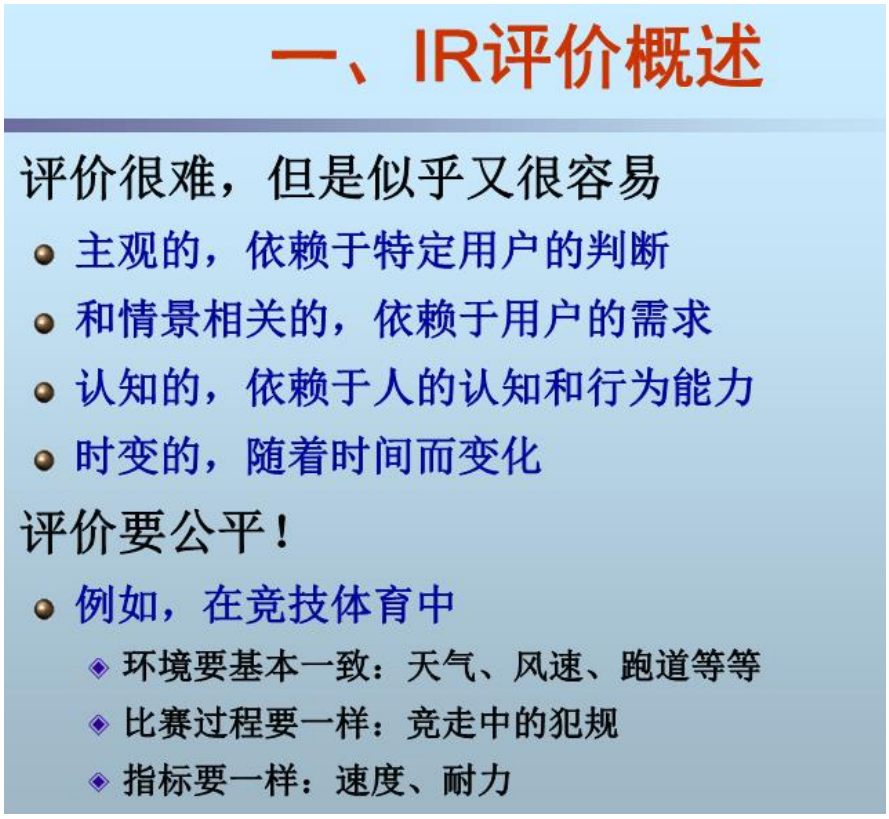
9. HITS

三、HITS

- **PageRank算法中对于向外链接的权值贡献是平均的，也就是不考虑不同链接的重要性。而WEB的链接具有以下特征：**
  - **1.有些链接具有注释性，也有些链接是起导航或广告作用。有注释性的链接才用于权威判断。**
  - **2.基于商业或竞争因素考虑，很少有WEB网页指向其竞争领域的权威网页。**
  - **3.权威网页很少具有显式的描述，比如Google主页不会明确给出WEB搜索引擎之类的描述信息。**
- 可见平均的分布权值不符合链接的实际情况

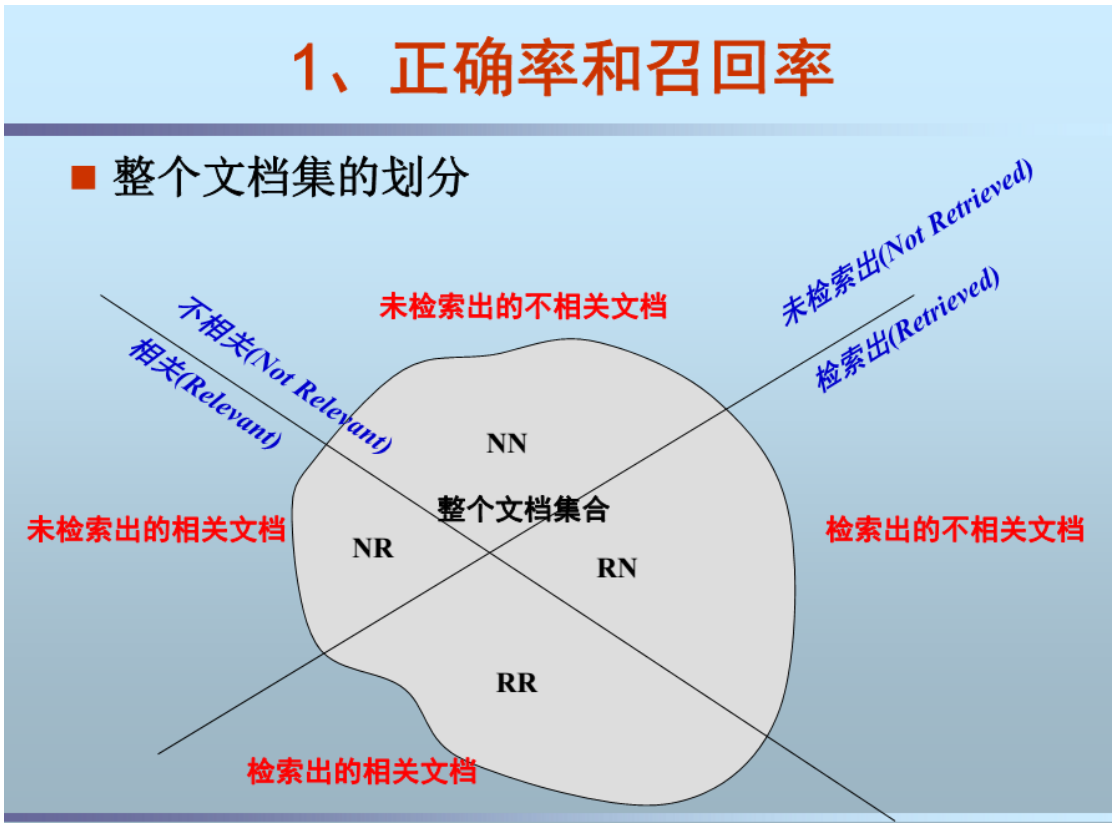
七． Evaluation

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



正确率(Precision)

- RR/(RR + RN)
- 返回的结果中真正相关结果的比率，也称为查准率，P∈ [0,1]

召回率(Recall)

- RR/(RR + NR)
- 返回的相关结果数占实际相关结果总数的比率，也称为查全率，R∈ [0,1]

两个指标分别度量检索效果的某个方面，忽略任何一个方面都有失偏颇。两个极端情况：返回有把握的1篇，P=100%，但R极低；全部文档都返回，R=1，但P极低

2、F-measure

F值(F-measure)：召回率R和正确率P的调和平均值，if P=0 or R=0, then F=0, else 采用下式计算：

$$F = \frac{2}{\frac{1}{P}+\frac{1}{R}} = \frac{2PR}{P+R} \quad (P \neq 0, R \neq 0)$$

更一般的情况——F_β：表示召回率的重要程度是正确率的β(>=0)倍，β>1更重视召回率，β<1更重视正确率。F值即β=1时的F_1值。F_2值（更重视召回率）和$F_{0.5}$值（更重视准确率）也是常用的指标值

$$F_\beta = \frac{(1+\beta^2)PR}{\beta^2 P + R} \quad (P \neq 0, R \neq 0)$$

4. P@N、R@Precision、AP 的定义

3、P@N

■ 举例

系统&查询	1	2	3	4	5
系统1，查询1	d3 √	d6 √	d8	d10	d11
系统1，查询2	d1 √	d4	d7	d11	d13 √
系统2，查询1	d6 √	d7	d2	d9 √	
系统2，查询2	d1 √	d2 √	d4	d13 √	d14

查询1的标准答案集合为 {d3,d4,d6,d9}
查询2的标准答案集合为 {d1,d2,d13}

系统1查询1: P@2=1, P@5=2/5;　　系统1查询2: P@2=1/2, P@5=2/5;

系统2查询1: P@2=1/2, P@5=2/5;　　系统2查询2: P@2=1, P@5=3/5

4、R-Precision

R-Precision

- 检索结果中，在所有相关文档总数位置上的准确率
- 如某个查询的相关文档总数为80，则计算检索结果中在前80篇文档的正确率。

系统&查询	1	2	3	4	5
系统1，查询1	d3 √	d6 √	d8	d10	d11
系统1，查询2	d1 √	d4	d7	d11	d13 √
系统2，查询1	d6 √	d7	d2	d9 √	
系统2，查询2	d1 √	d2 √	d4	d13 √	d14

查询1的标准答案集合为 {d3,d4,d6,d9}　查询2的标准答案集合为 {d1,d2,d13}

系统1查询1: R-Precision=2/4;　系统1查询2: R-Precision=1/3;

系统2查询1: R-Precision=2/4;　系统2查询2: R-Precision=2/3;

5. MAP、MRR

6、MAP

MAP(Mean AP)

- 对所有查询的AP求算术平均
- 反映在全部查询上的检索效果

例如：假设有一个检索系统

- 对查询1返回4个相关网页，其rank分别为1, 2, 4, 7
- 对查询2返回3个相关网页，其rank分别为1, 3, 5
- 查询1共有4个相关文档，查询2共有5个相关文档

查询1：AP ＝ (1/1+2/2+3/4+4/7)/4＝0.83
查询2：AP ＝ (1/1+2/3+3/5+0+0)/5＝0.45
MAP ＝ (0.83+0.45)/2 ＝ 0.64

7、MRR

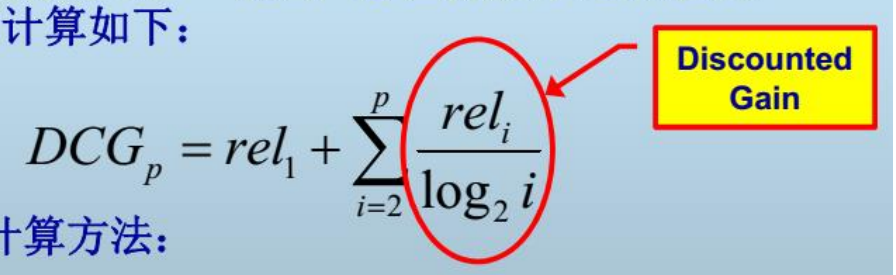
- **MRR(Mean Reciprocal Rank)**
  - 对于某些IR系统(如问答系统或主页发现系统)，只关心第一个标准答案返回的位置(Rank)，越前越好，这个位置的倒数称为RR，对问题集合求平均，则得到MRR
- 例如
  - 两个问题，系统对第一个问题返回的标准答案Rank是2，对第二个问题返回的标准答案的Rank是4
  - 则系统的MRR＝ (1/2+1/4)/2＝3/8
  - 意味着平均在第8/3个位置处找到相关文档

6. NDCG

8、NDCG

DCG(Discounted Cumulative Gain)

- 基本思想: 若搜索算法把相关度高的文档排在后面，则应该给予惩罚。一般用log 函数表示这种惩罚。DCG 的计算如下：



$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

- 另一种计算方法：

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(1+i)}$$

更强调排在前面的相关文档的重要性（指数）

8、NDCG

- DCG计算例子
  - 相关度0－3，10个文档的得分如下:
    - 3, 2, 3, 0, 0, 1, 2, 2, 3, 0
  - discounted gain:
    - ＝ 3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0
      ＝ 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0
  - DCG＝3＋∑ discounted gain:
    - 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

8、NDCG

- DCG的值与具体查询有关，和结果列表的长度有关，不利于检索系统之间的对比
  - 不同query的搜索结果有多有少，所以不同query的DCG值就没有办法来做对比
  - 例如，DCG_5＝6.89，DCG_{10}＝9.61
- NDCG (Normalized DCG)：对DCG进行规范化
  - 把检索结果按相关度从大到小排序得到一个理想的输出序列
  - 计算此理想序列的DCG, 得到在位置*p* 的ideal DCG（IDCG)
  - 然后以位置*p*的DCG_p与$IDCG_p$比值作为评价指标

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

8、NDCG

- NDCG计算示例
  - 沿用前面例子：相关度0－3，10个文档的得分如下:
    - 3, 2, 3, 0, 0, 1, 2, 2, 3, 0
  - 理想的输出结果序列: 3, 3, 3, 2, 2, 2, 1, 0, 0, 0
- ideal DCG (IDCG):
  - 3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 10.88, 10.88, 10.88
- DCG＝3＋Σ discounted gain:
  - 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61
- NDCG: ($NDCG_i = DCG_i / IDCG_i$)

  1, 0.83, 0.87, 0.76, 0.71, 0.69, 0.73, 0.8, 0.88, 0.88

  可以看到任何查询结果位置*p*的NDCG值都规范化为 ≤ 1 的值

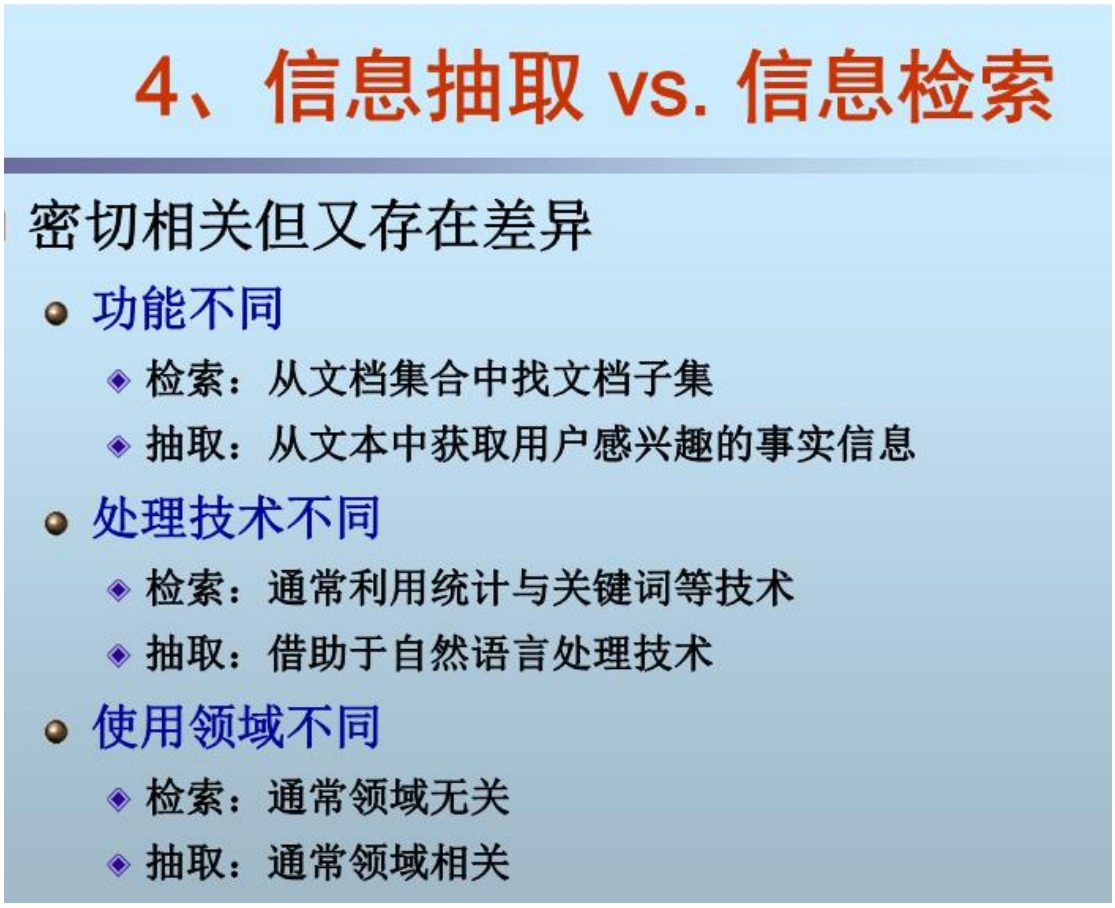
PART 2：Web Information Extraction

一、 Named Entity Recognition

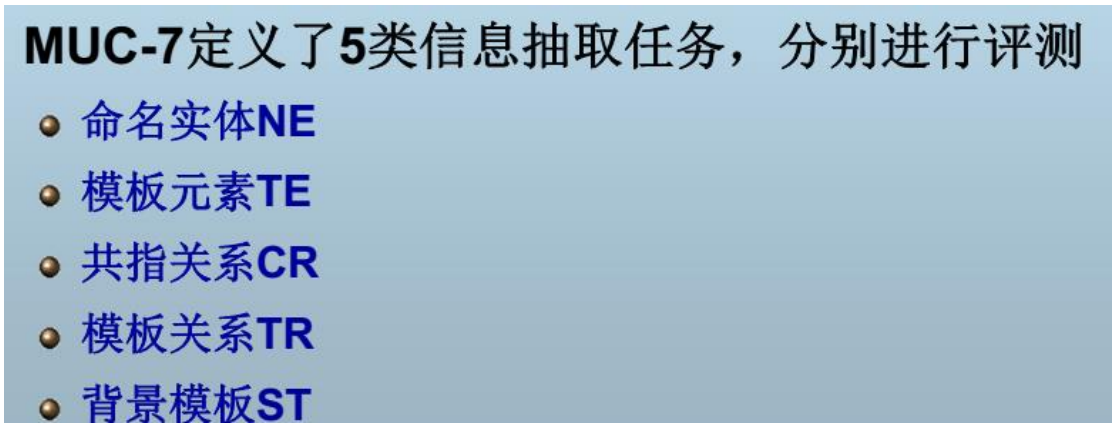
1. 信息抽取（IE）的概念以及与 IR 的关系

1、信息抽取含义

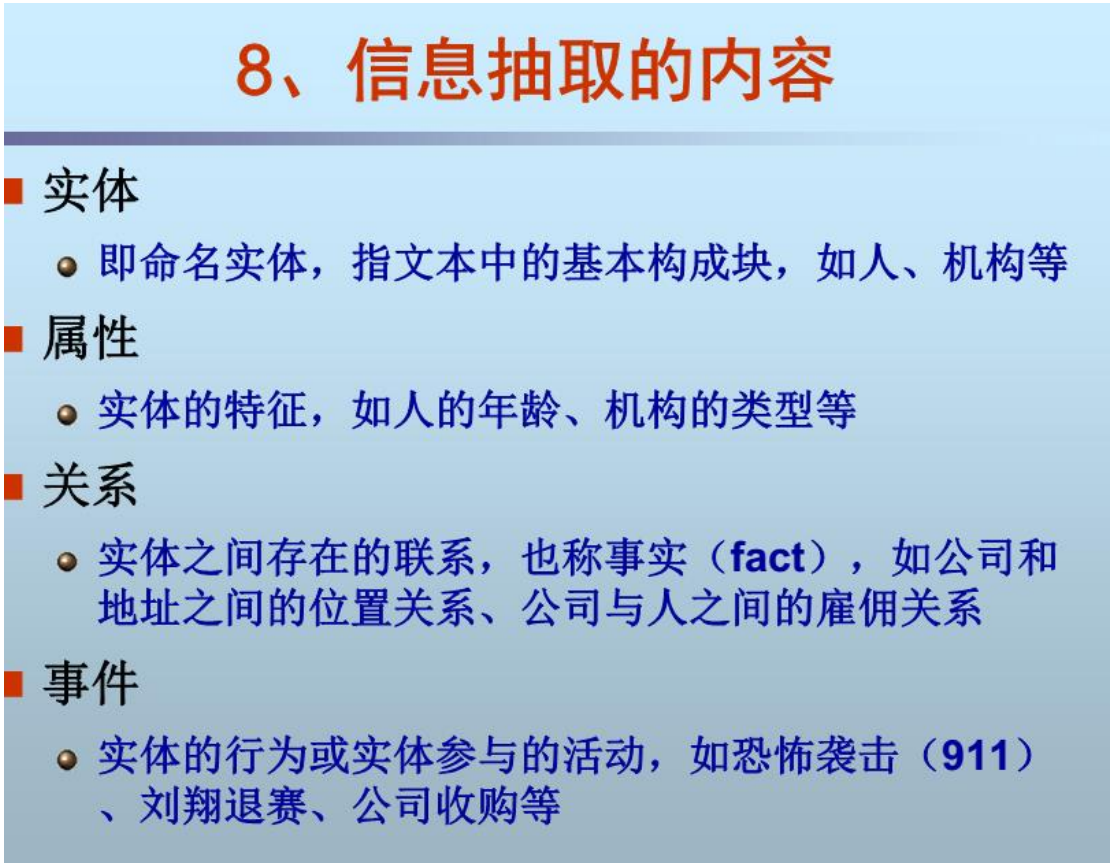
- 从一段文本中抽取指定的事件、事实等信息，形成结构化的数据
  - 从文本中抽取用户感兴趣的事件、实体和关系
  - 被抽取的信息以结构化的形式描述
  - 为情报分析、检测、比价购物、自动文摘、文本分类等各种应用提供服务



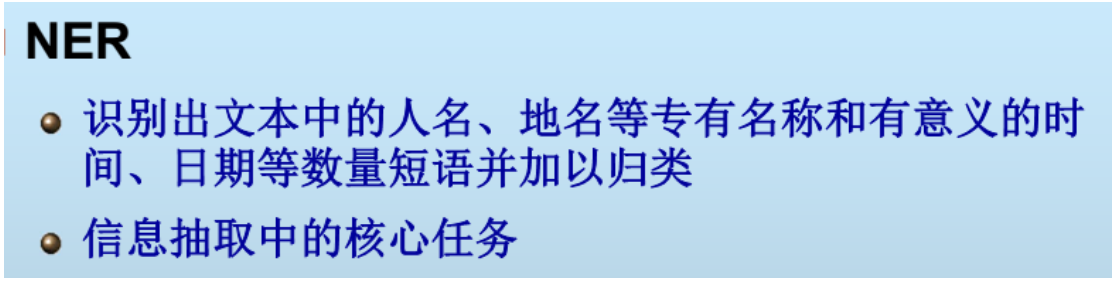
2. MUC－7 定义的信息抽取任务



3. 信息抽取的内容

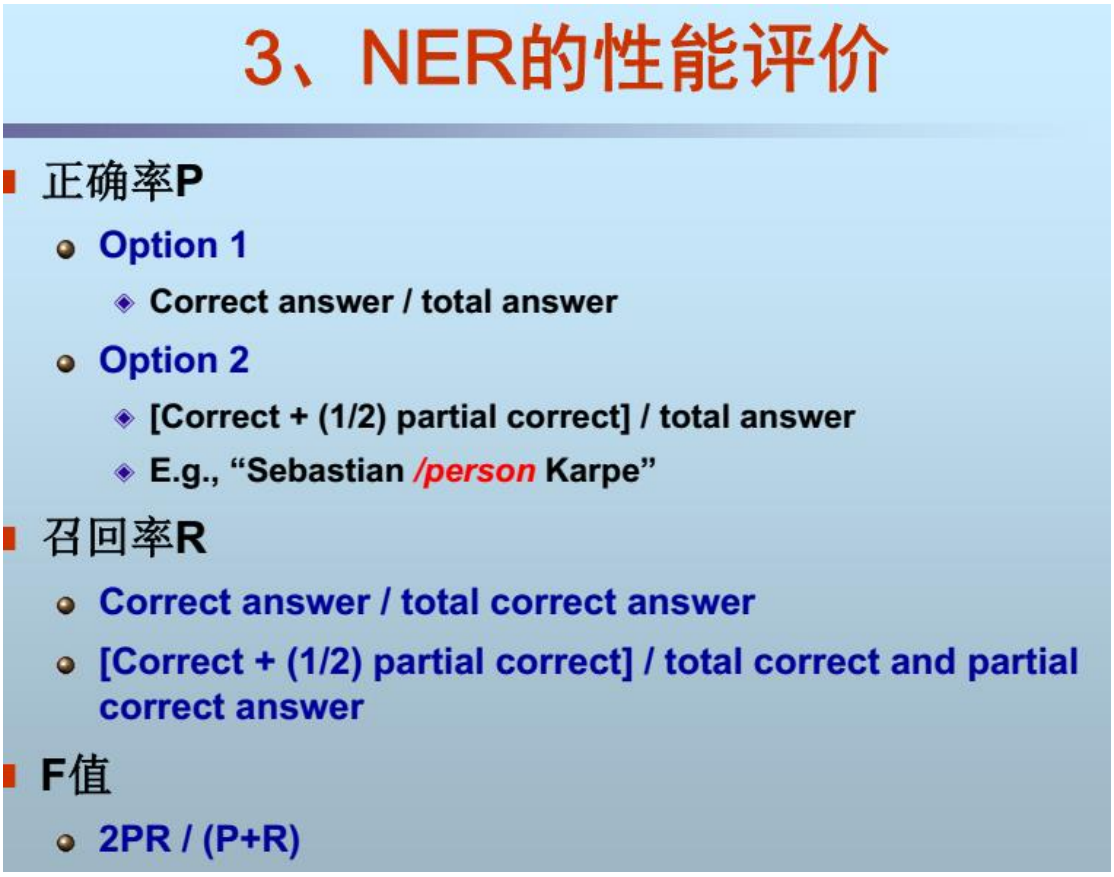

8、信息抽取的内容
实体
即命名实体，指文本中的基本构成块，如人、机构等
属性
实体的特征，如人的年龄、机构的类型等
关系
实体之间存在的联系，也称事实（fact），如公司和地址之间的位置关系、公司与人之间的雇佣关系
事件
实体的行为或实体参与的活动，如恐怖袭击（911）、刘翔退赛、公司收购等

4. NER 的概念与难点


NER
识别出文本中的人名、地名等专有名称和有意义的时间、日期等数量短语并加以归类
信息抽取中的核心任务

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

关系抽取

- 从文本中识别出两个实体或多个实体之间存在的事实上的关系。

> …在文本中检测实体之间语义关系的一种技术…
>
> *摘录自 维基百科*

- 例如

Bill Gates worked at Microsoft.

↓

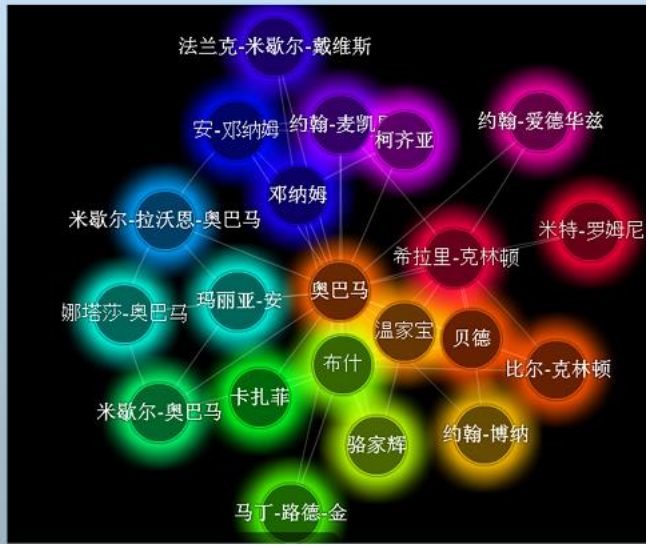
beEmployee(Bill Gates, Microsoft)

- 1997年，MUC-7上首次引入了关系抽取任务（Template Relation）

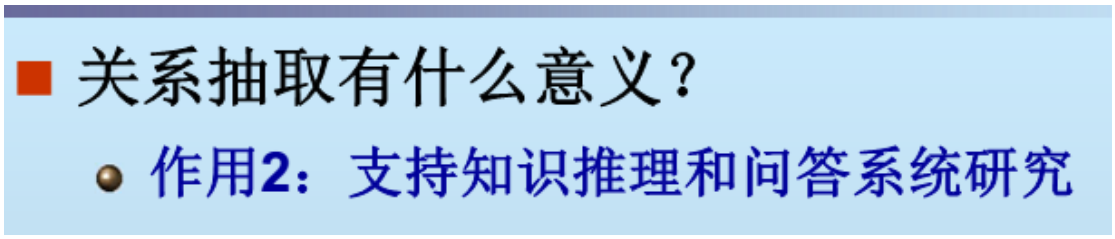
一、关系抽取概念

- 关系抽取有什么意义？
  - 作用1：提高搜索引擎发现知识的能力

http://renlifang.msra.cn/

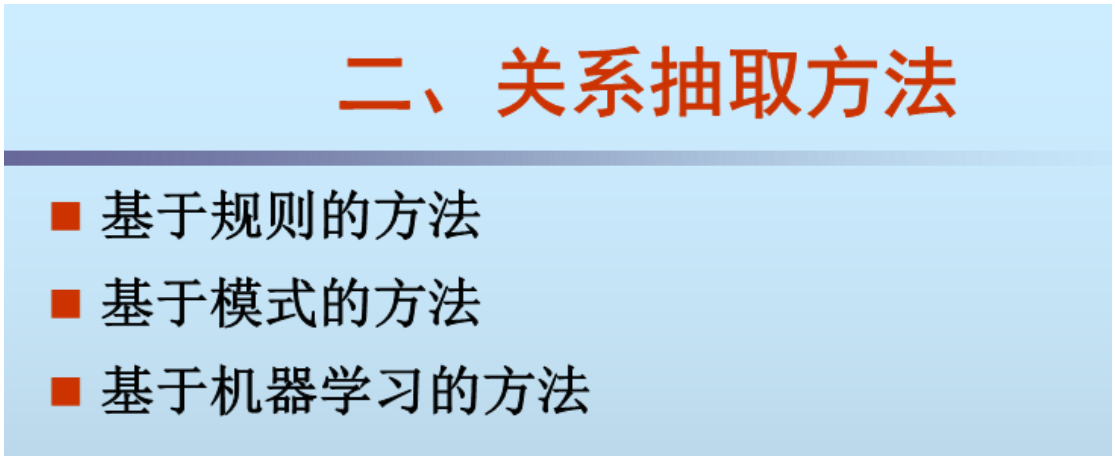



一、关系抽取概念
关系抽取有什么意义？
作用2：广泛应用于各种知识库的构建
DBpedia
YAGO
http://dbpedia.org/
http://www.mpi-inf.mpg.de/yago-naga/ya
超过三百四十万个实体
以及十亿个实体关系
超过两百万个实体
以及两千万实体关系


关系抽取有什么意义？
作用2：支持知识推理和问答系统研究

2. 关系的表示方法

3. 关系抽取的常用方法



PART 3: Web Mining

一. 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？

Discovering useful information from the World-Wide Web and its usage patterns

Web Mining Topics

Content mining

Structure mining

Usage mining

Web Content Mining

Definition: Web content mining is the process of extracting useful information from the contents of Web documents.

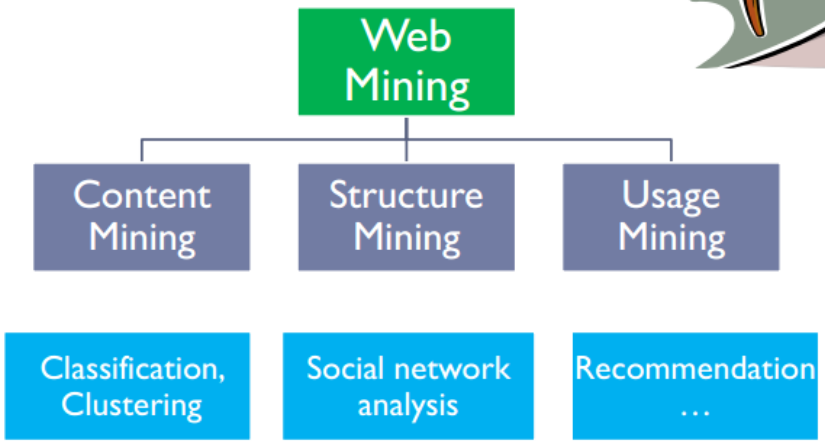
- Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables.

- Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR—信息检索) and Natural Language Processing (NLP—自然语言处理).

Web Structure Mining

- Generate *structural summary* about the Web site and Web page
  - Hierarchy of hyperlinks in the website and its structure.
- Finding information about web pages
  - Retrieving information about the relevance and the quality of the web page.
  - Finding the authoritative（权威性，可信度） on the topic and content.
- Inference on hyperlinks
  - The web page contains not only information but also hyperlinks, which contains huge amount of annotation.
  - Hyperlink identifies author's endorsement of the other web page

Roadmap



Note: Helpful to combine usage with content and structure

Web Usage Mining

Navigation Patterns

- Examples:

  70% of users who accessed /company/product2 did so by starting at /company and proceeding through /company/new, /company/products and company/product1

  80% of users who accessed the site started from /company/products

  65% of users left the site after four or less page references

二． Web Content Mining

数据(Data)

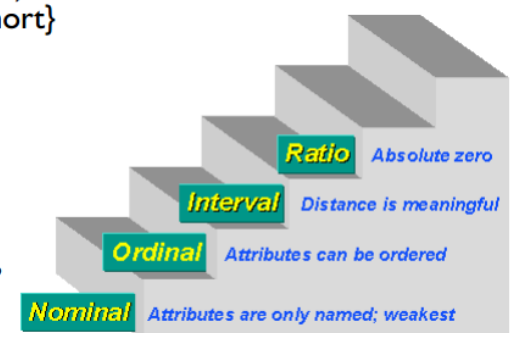
1. 概念：数据对象(Objects)，属性(Attributes)，维度(Dimensions)，特征(features)

Attributes

- Attribute (or dimensions, features, variables): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*

Types of Attributes

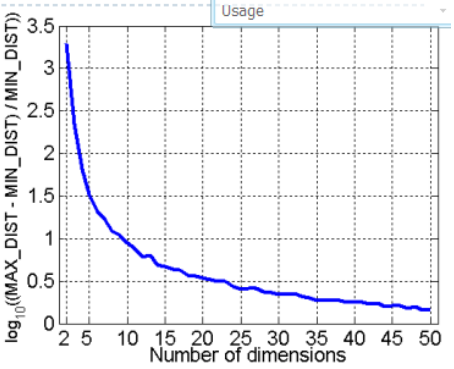
- There are different types of attributes
  - Nominal （标称）
    - Examples: ID numbers, eye color, zip codes
  - Ordinal （序数）
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval （区间）
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio （比例）
    - Examples: temperature in Kelvin, length, time, counts



2. 高维诅咒(Curse of dimensionality)现象。

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies
- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful
- If N_1 = 100 represents a dense sample for a single input problem, then $N_{10} = 100^{10}$ is the sample size required for the same sampling density with dimension 10.
- The proportion of a hypersphere with radius *r* and dimension *d*, to that of a hyercube with sides of length 2*r* and dimension *d* converges to 0 as *d* goes to infinity — nearly all of the high-dimensional space is "far away" from the center



- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points

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。

Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Aggregation
☐ Sampling
☐ Dimensionality Reduction
☐ Feature subset selection
☐ Feature creation
☐ Discretization and Binarization
☐ Attribute TransformationAggregation
☐ Sampling
☐ Dimensionality Reduction
☐ Feature subset selection
☐ Feature creation
☐ Discretization and Binarization
☐ Attribute Transformation

聚集的
☐ 抽样
☐ 降维
☐ 特征子集选择
☐ 特征创建
☐ 离散化和二值化
☐ 属性 transformationaggregation
☐ 抽样
☐ 降维
☐ 特征子集选择
☐ 特征创建
☐ 离散化和二值化
☐ 属性变换
☐ 双语对照

Sampling

- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.

Sampling

- The key principle for effective sampling is the following:
  - Using a sample will work almost as well as using the entire data sets, if the sample is representative
  - A sample is representative if it has approximately the same property (of interest) as the original set of data

Usage

Feature Subset Selection

- Another way to reduce dimensionality of data
- Redundant features
  - Duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - Contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA
- Many techniques developed, especially for classification

Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

- Three general methodologies:
  - Feature extraction

    Example: extracting edges from images
  - Feature construction

    Example: dividing mass by volume to get density
  - Mapping data to new space

    Example: Fourier and wavelet analysis

Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

- Techniques
  - Principal Components Analysis (PCA)
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

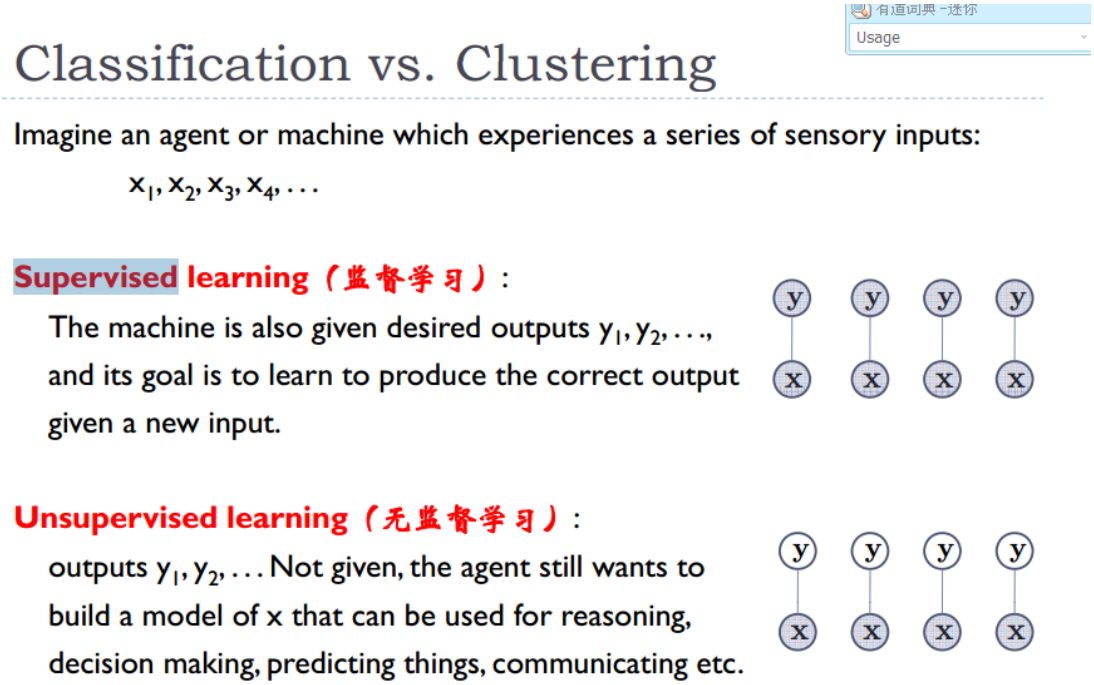
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More "stable" data
    - Aggregated data tends to have less variability

降维技术：主成分分析、奇异值分解、监督和非线性技术

分类(Classification)

手工、自动文件分类、文件标签分配功能监督学习 Supervised learning of a document-label assignment function

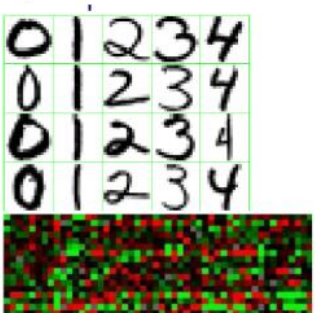
4. 监督学习(Supervised learning)与无监督学习(Unsupervised learning)的关系与区别。



5. 分类(Classification)的基本原理。

Classification

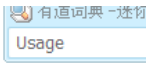
- We are given a set of N observations $\{(\mathbf{x}_i, y_i)\}_{i=1..N}$
  - Issue: how to represent data: text, image, video…
- Need to map x $\in \mathcal{X}$ to a label y $\in \mathcal{Y}$
  - We want to know how to build classification functions ("classifiers").
- Examples:



digits recognition;
$\mathcal{Y} = \{0, \ldots, 9\}$

6. 数据的向量表示(Vector space representation)

Vector Space Representation



Web document classification

Each document is a vector, one component for each term (=word).

	Doc 1	Doc 2	Doc 3	…
Word 1	3	0	0	…
Word 2	0	8	1	…
Word 3	12	1	10	…
…	0	1	3	…
…	0	0	0	…

High-dimensional vector space:

- Terms are axes, 10,000+ dimensions, or even 100,000+
- Docs are vectors in this space

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。

3、5、欧式距离和 Hamming distance

Nearest-Neighbor Learning Algorithm

Learning is just storing the representations of the training examples in *D*.

Testing instance *x*:

- Compute similarity between *x and all examples in D.*
- Assign *x the category of the majority of the k most similar examples in D.*

Also called:

- Case-based learning（基于实例的学习）
- Memory-based learning
- Lazy learning

k Nearest-Neighbor

- Using only the closest example to determine the categorization is subject to errors due to:
  - A single atypical example.
  - Noise (i.e. error) in the category label of a single training example.
- More robust alternative is to find the *k* most-similar examples and return the majority category of these *k* examples.
- Value of *k* is typically odd to avoid ties; 3 and 5 are most common.

8. 熟练掌握最小二乘算法——推导过程，闭式解，规范化之后的求解推导。

Least Squares Classification

Least squares loss function:

$$L_2(h(\mathbf{x}), y) = (y - \mathbf{w}^\top \mathbf{x} - b)^2$$

The goal:

to learn a classifier $h(\mathbf{x}) = \text{sign}(\mathbf{w}^\mathsf{T}\mathbf{x} + b)$ to minimize the least squares loss

$$\begin{aligned} Loss &= \min_{\mathbf{w},b} \sum_i L_2(h(\mathbf{x}_i), y_i) \\ &= \min_{\mathbf{w},b} \sum_i (y_i - \mathbf{w}^\top \mathbf{x}_i - b)^2 \end{aligned}$$

Solving Least Squares Classification

Let

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ \vdots & & \vdots & \\ 1 & x_{N1} & \cdots & x_{Nd} \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} b \\ \vdots \\ w_d \end{bmatrix}$$

$$\begin{aligned} Loss = \min_{\mathbf{w}} (\mathbf{y} - X\mathbf{w})^2 &= \min_{\mathbf{w}} (X\mathbf{w} - \mathbf{y})^2 \\ &= \min_{\mathbf{w}} (X\mathbf{w} - \mathbf{y})^\top (X\mathbf{w} - \mathbf{y}) \end{aligned}$$

Solving for w

$$\frac{\partial Loss}{\partial \mathbf{w}} = 2(X\mathbf{w} - \mathbf{y})^\top X = 0$$

$$X^\top X\mathbf{w} - X^\top \mathbf{y} = 0$$

$$\mathbf{w}^* = (X^\top X)^{-1} X^\top \mathbf{y}$$

Note: $d(\mathbf{Ax}+\mathbf{b})^T\mathbf{C}(\mathbf{Dx}+\mathbf{e}) = ((\mathbf{Ax}+\mathbf{b})^T\mathbf{CD} + (\mathbf{Dx}+\mathbf{e})^T\mathbf{C}^T\mathbf{A})\, d\mathbf{x}$

$d(\mathbf{Ax}+\mathbf{b})^T(\mathbf{Ax}+\mathbf{b}) = (2(\mathbf{Ax}+\mathbf{b})^T\mathbf{A})\, d\mathbf{x}$

- $X^+ = (X^\top X)^{-1} X^\top$ is called the *Moore-Penrose* pseudoinverse（伪逆） of X
- Least squares classification in Matlab

```
% X(i: ,) is the i-th example, y(i) is the i-th label
wLSQ = pinv([ones(size(X, 1), 1) X])*y;
```

- Prediction for $\mathbf{x}_0$

$$\hat{y} = \text{sign}\left(\mathbf{w}^{*\top} \begin{bmatrix} 1 \\ \mathbf{x}_0 \end{bmatrix}\right) = \text{sign}\left(\mathbf{y}^\top X^{+\top} \begin{bmatrix} 1 \\ \mathbf{x}_0 \end{bmatrix}\right)$$

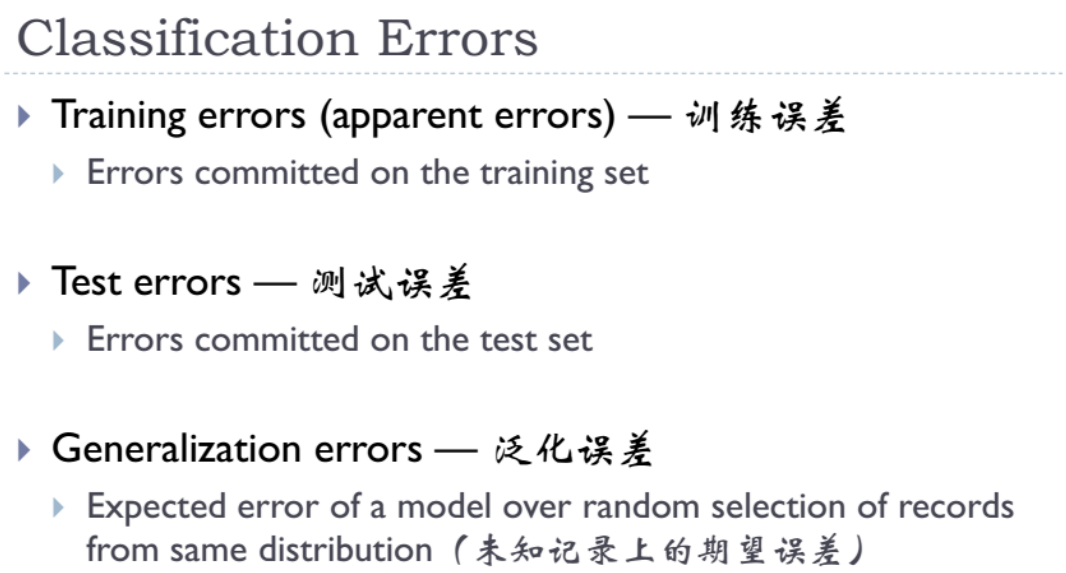
45

9. 过拟合现象出现的原因。

Model Overfitting

- Due to noise
- Due to insufficient examples

10. 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)

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

三． 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。

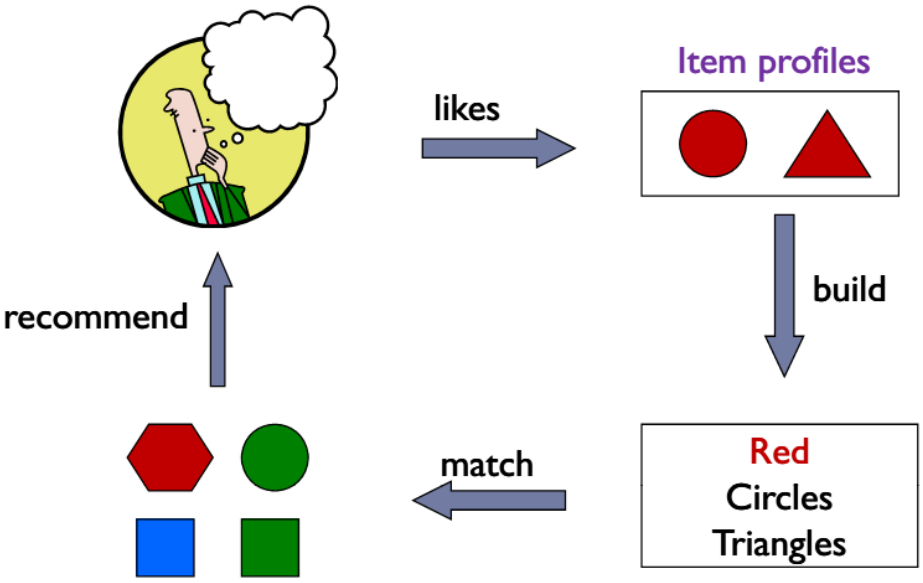
影响力分析(Influence)

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． Web Recommendation

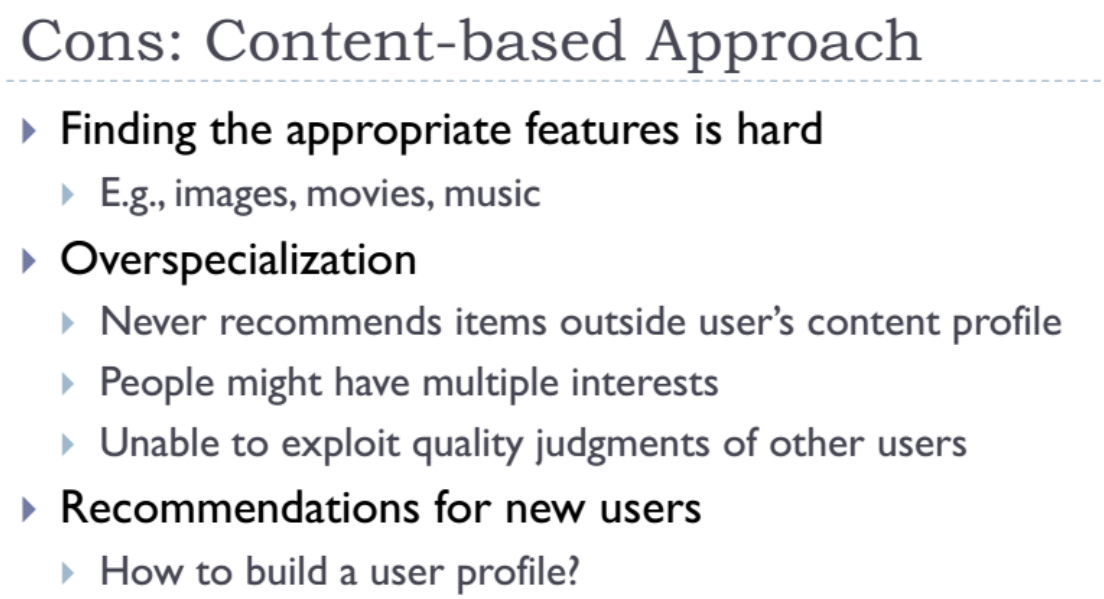
1. 推荐系统基本模型以及一般工作流程。
2. 基于内容的推荐算法流程及优缺点

Plan of Action



Pros: Content-based Approach

- No need for data on other users
  - No cold-start or sparsity problems
- Able to recommend users with unique tastes
- Able to recommend new and unpopular items
  - No first-rater problem
- Able to provide explanations
  - Can provide explanations of recommended items by listing content-features that caused an item to be recommended



3. 协同过滤推荐算法流程及优缺点

部分未完成。

考试题目：

1.十个判断题 20

2．IR 系统的正确率、召回率、F、MAP 等解释，为何需要同时用正确率和召回率。

3．倒排索引

4.tf-idf、余弦

5.K-means 是否一定收敛，并证明。

6.normal。。。 cut 1、2、3、4、5、6

7.最小二乘加权 r 的推导

Web Information Processing and Applications

Instructor

Jin Pei-Quan（金培权）
Email: jpq@ustc.edu.cn

Xu Lin-Li （徐林莉）
Email: linlixu@ustc.edu.cn

Teaching Assistants

林盛, Ph.D. student
Phone: 13485728758
Email: linsh@mail.ustc.edu.cn
Room: 1610, 科技实验楼西楼

于永波, Master Student
Phone: 13865979122
Email: yyb2012@maiul.ustc.edu.cn
Room: 1610, 科技实验楼西楼

Lectures

Time: Class 6 to 8

Classroom: 3C221 (West Campus)

Textbook

W. Bruce Croft, Donald Melzler, Trvor Strohman, ***Search Engines: Information Retrieval in Practice***, Pearson Press, 2010

(中文版：刘挺, 等 译, 搜索引擎：信息检索实践, 机械工业出版社, 2012)

Soumen Chakrabarti, ***Mining the web: discovering knowledge from hypertext data***, 人民邮电出版社, 2009

Anand Rajaraman, Jeffrey David Ullman. 王斌 译, ***Mining of Massive***

Datasets. 人民邮电出版社, 2012

References

Christopher D. Manning, Prabhakar Raghavanm, Hinrich Schütze, ***An Introduction to Information Retrieval***, Cambridge University Press, 2008 (中文版：王斌 译, 信息检索导论, 人民邮电出版社, 2010)

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, ***Modern Informatio Retrieval,*** Addison Wesley Longman Publishing Co. Inc., 1999

Bing Liu, ***Web Data Mining*** (2nd Edition), Springer, 2011

Some state-of-the-art papers from SIGIR, CIKM, WWW, etc.

Assignments

Some homework assignments. **POLICY: all assignments should be completed and submitted in one week, i.e. before the beginning of next class. Late assignment submissions will be penalized 20% points.**

Examination

One final test, scheduled to be taken at the end of the course.

Grading

Homework: 20%
Lab: 20% **[** Lab #1 Description**.** NEW Lab #2 Description**. Lab time: 18:30-21:30, Tuesday, start from 8 October. Lab site: 517, E3 Building]**
Final: 60%

Course Notes

No.	Date	Contents	Homework	Chapters Reading
1	9.3	**Introduction to Web Information Processing**		**Chp.1-2**
2	9.10	**Web Crawling** (NEW	**homework**	**Chp.3**

		updated)		
3	9.17	**Text Processing**	**homework**	**Chp.4**
4	9.24	**Indexing & Lab #1 Description**	**homework**	**Chp.5**
5	10.1	(***National Day***)	NEW **Lab #1 Lab time: 18:30-21:30, Monday and Tuesday. Location: 517, E3 Building**	
6	10.8	**Queries**	**homework**	**Chp.6**
7	10.15	**Ranking**	**homework**	**Chp.7**
8	10.22	**Evaluation**		**Chp.8**
9	10.29	**Named Entity Recognition**	**homework**	
10	11.5	**Relation Extraction**		
11	11.12	**Web Mining: Introduction**		
12	11.19	**Web Mining: Data**		
13	11.26	**Web Mining: Classification**	**homework**	
14	12.3	**Web Mining: Clustering**		
15	12.10	**Web Mining: Social Network Analysis I**	**homework**	
16	12.17	**Web Mining: Social Network Analysis II**	NEW **Lab #2 Description**	
17	12.24	**Web Mining: Social Network**		

		Analysis III		
18	12.31	**Web Mining: Recommendation**		
19	1.3 (Friday)	**Review & Q/A**	**[19:00 PM, 3C221]**	
20	1.9	**Final Exam**	NEW **8:30AM-10:30AM Room 3C121 & 3C122**	