第一讲 概述

1.信息检索主要步骤：爬行和抓取，建立索引，搜索词处理，排序

2.搜索引擎评价：覆盖面，更新周期，响应速度，排序结果是否满足用户的查询要求

3.SEO：搜索引擎优化，再好的搜索引擎也不能与人相比，搜索引擎常常并不能正确返回最相关最有用的信息，因此需要优化

4.网络爬虫：按照一定的规则，自动地抓取万维网信息的程序或者脚本

5.建立索引：对爬虫抓取的页面进行解析处理，并以巨大的表格形式存入数据库

6.UN：用户需求，用户需要获得的信息

7.文档：检索的对象，可以是文本，也可以是图像、视频、语音等多媒体文档

8.文档集合：所有待检索的文档构成的集合

9.相关、相关度、相似度：用于描述检索关键词与返回结果之间的联系，取决于用户的判断，是一个主观概念

10.定义相关性的两个角度：

系统角度：系统输出结果，用户是信息的接受者

用户角度：观察用户对检索结果的反应，是系统输出向用户需求的投射

11.信息搜索：给定一个查询Q，从文档集合C中计算每篇文档D与Q的相关度并排序

12.搜索引擎核心技术：网页爬取，网页预处理，文本处理(分词，词干提取，去停用词)，建立索引，查询，Rank，用户反馈

13.网页预处理：去除网页结构，去除tag，保留url，保留文本

14.文本处理：对查询和文本进行的预处理操作

15.建立文本索引：对经过文本处理后的文本进行进一步处理，得到文本的内部表示，通常基于索引项来表示

16.搜索：从文本中查找包含查询中索引项的文本

17.排序：对搜索出的文本按照某种方式来计算相关度

18.IR的两种模式：拉和推

19.Lucene：由Apache软件基金会支持和提供，是一个开放源代码的全文检索引擎工具包

提供了完整的查询引擎和索引引擎，部分文本分析引擎。主要分为索引模块和检索模块

org.apache.Lucene.search 搜索入口 index 索引入口 analysis 语言分析器

queryParser 查询分析器 document 存储结构 store 底层IO存储结构 util 数据结构

20.Solr：Solr的底层核心技术是使用Lucene来实现的

Solr和Lucene的本质区别有三点：搜索服务器，企业级和管理

Lucene本质是搜索库，不是独立的应用程序，而Solr是应用程序

Lucene专注于搜索底层的建设，而Solr专注于企业应用

Lucene不负责支撑搜索服务所必须的管理，而Solr负责

Solr是Lucene面向企业搜索应用的扩展

第二讲 网络爬虫技术

1.网络爬虫定义：是一种自动获取网页内容的程序或脚本，是搜索引擎的重要组成部分。通过HTML源码解析来获得想要的东西

2.爬虫过程：从一个或若干个初始页面的url开始，在抓取网页的过程中，消除噪音，区分url是否抓取，再根据一定的搜索策略选择新的url放入队列，直到满足系统一定的停止条件。注意调度，控制抓取间隔。

3.构建网络爬虫的工程要点：调度策略(BFS,DFS,最佳优先)，访问标记(建立一个散列，存放访问过的每一个网址)

4.爬虫服务器分布化：问题1哈希表太大一个服务器装不下；问题2每台服务器开始下载前和下载完成后都要维护哈希表，存储哈希表的通信会影响效率

问题2解决方法：明确每台服务器的分工，看到url就知道应该让哪一台服务器去执行；批量处理，减少通信的次数

5.常见的开源爬虫：Heritrix，Nutch，pyspider，scrapy

6.Nutch：是一个开源JAVA实现的搜索引擎。提供了运行自己的搜索引擎所需的全部工具。包括全文搜索和Web爬虫。

主要特点：多线程，宽度优先，遵从robots.txt，采用socket连接并getInputStream，边爬取边解析(获取一个页面后对其进行解析，存储原始页面和解析后的文档)，页面打分

第三讲 网页分析技术

1.正则表达式定义：对字符串操作的一种逻辑公式，用事先定义好的一些特定字符的组合，组成"规则字符串"。用这个"规则字符串"来表达对字符串的一种过滤逻辑。

2.正则表达式作用：用来检索、替换负责某种规则(模式)的文本

3.正则表达式匹配特点：匹配速度快；表达能力弱，只具有正规文法的表示能力；在网页内容信噪比要求不高的情况下可以使用，否则检索结果存在无用数据

4.DOM：文档对象模型，将XML或HTML转化为树状结构的元素，所有元素及他们的文字和属性都可以通过对DOM树来操作或访问

5.HTML DOM树特点：在解析HTML时速度较慢，其表达能力相当于上下文无关文法；在网页自动分类等需要网页去噪处理的情况时可使用。

6.HTML和正则表达式的差异比较

7.举出一个HTML解析的例子(Beauteful soup)

8.元搜索引擎的定义：通过一个统一的用户界面帮助用户在多个搜索引擎中选择和利用合适的搜索引擎来实现检索操作，是对于多种检索工具的全局控制机制

9.网页反爬虫措施：

robots协议：网页通过robots协议来规范爬虫哪些网页可以爬取哪些不可以

解决方法：当爬虫访问一个站点时，应该先去检查是否存在robots协议，若存在，则按照文件中的内容来规范自己的访问范围

IP屏蔽：同一IP频繁访问后会被封

解决方法：连接代理服务器，多IP并行，增大爬取时间间隔

访问限制：交互登陆(提供用户名和密码)，JS渲染，动态网页(数据在后台服务器)

解决方法：通过PhantomJS或Selenium来模拟浏览器操作，并通过HTTP分析工具来分析HTTP传递的口令

验证码

解决方法：用脚本或人工对其图片进行爬虫遍历(得到12306的图片库)，将所有图片保存后(分割验证码图像)与关键字进行对比并关联入库(置入百度识图API函数并返回识图结果)

第四讲 词项词典

1.词条化：将给定的字符序列拆分成一系列子序列的过程，其中每一个子序列称之为一个"词条"token

2.词条和词项的区别：字符序列拆分后的每一个子序列是token，所有token生成的词项词典中的每一项被称为词项

3.消除停用词的意义：停用词消除可以减少term的个数；可以缩小搜索范围；可以提高搜索的效率；有效提高关键词密度

4.消除停用词的缺点：一些有意义的词可能会被消除，比如to be or not to be，的士

5.消除停用词的方法：语法剔除(冠词，介词，代词)；基于文档频率(消除频率高的无意义词)；基于停用词表

6.Stemming：词干还原，指去除单词两端词缀的启发式过程，能够提高召回率，但是会降低准确率。

7.常用的词干还原算法：Porter(英文处理)，

8.Lemmatization：词性归并，利用词汇表和词形分析来减少屈折变化的形式，将其转化为基本形式。

9.Stemming与Lemmatization的区别：代表意义不同，stemming指的是去除单词两端词缀的启发式过程，Lemmatization指的是利用词汇表和词形分析来减少屈折变化的形式从而返回词的原型的过程。Stemming通常会将多个派生相关词合并到一起，而Lemmatization会将同一词元的不同曲折形式进行合并。Stemming和Lemmatization都体现了不同语言之间的差异性。

第五讲 中文分词

1.分词算法：基于字符串匹配的分词方法(遇到字典里有的词就标识出来，遇到复合词就找最大匹配，遇到不认识的词就切割成单个字)；基于理解的分词方法；基于统计的分词方法(字与字相邻共现的频率能够较好地反映词的可信度)

2.统计语言模型：是一个单词序列可能的分布

3.n-gram模型：第n个词的出现只与前面n-1个词相关，与其他任何词都不相关，整句的概率就是各个词出现概率的乘积

特点：简单有效；只考虑了词的位置关系，没有考虑词之间的相似度、词语法和词语意，存在数据稀疏的问题(造成零概率)

4.古德-图灵估计的主要思想：是一种实用的平滑算法

5.隐马尔可夫模型：用来描述一个含有隐含未知参数的马尔可夫过程。很重要，待看。

6.常用的开源分词软件：庖丁解牛，结巴分词

第六讲 信息检索模型和布尔检索

1.IR模型的四元组：D,Q,R(q,d),F，文档集合，查询集合，排序函数，框架(用于构建文档、查询和他们之间关系的模型)

2.基于内容的信息检索模型：集合论模型(布尔模型等)；代数模型(向量空间模型等)；概率模型(经典概率论模型等)

3.布尔检索的定义：是一种建立在经典集合论和布尔代数基础上的简单检索模型，可以用来处理布尔表达式形式的查询

词袋模型的概念

4.布尔检索模型的优点：查询简单，容易理解；通过使用复杂的布尔表达式，可以方便地控制查询结果；相当有效的实现方法；用户可以容易写出布尔表达式；通过扩展可以使其包含排序的功能

5.布尔检索模型的缺点：信息需求的能力表达不足，不能输出部分匹配的情况；无权重设计，无法排序；用户必须要学会用布尔表达式查询，一般情况下检索出的文档过多或过少；很难进行自动的相关反馈；对于大规模文档的检索太慢

6.词项文档矩阵避免占空间过大的方法：只记录1的位置

第七讲 倒排索引

1.倒排索引的概念：是搜索引擎的核心数据结构，主要包括词项词典和倒排记录表。

词项词典：对于每一个词项，存储所有包含这个词项的文档的一个列表

倒排记录表：其中每一个文档用一个序列号docID来表示

第八讲 向量空间模型

1.排序检索模型：系统根据文档与查询的相关度排序返回文档集合中的文档，分为布尔查询和自由文本查询两种方式。

2.排序检索的评分方法：

Jaccard系数：Jaccard(A,B) = |A∩B|/|A∪B| ，缺点：没有考虑词项频率，没有考虑罕见词和常见词的权重不同

3.词袋模型：不考虑词在文档中出现的位置的一种模型

4.词袋模型的改进：二元词，基于位置

5.tf-idf的概念：tf-idf是一种统计方法，用于评估一个词项对于其所在文件的重要程度。词项的重要性随在该文件中出现次数增多而增加，随该词项在语料库中出现次数的增多而减小

tf是词项频率(词项t在文档d中出现的次数 1+log10tf)，idf是逆文档频率(log10(N/df))，df是文档频率(出现词项的文档数目)

6.tf-idf的意义：tf-idf权重计算是排序检索的一种排序方法，相较于只考虑词项频率外，该方法还考虑了词项在整个文档集中的频率进行权重和评分计算，对常见词赋予低权重，对罕见词赋予高权重，使得检索排序结果更准确

7.向量空间模型的概念：将每篇文档表示成一个基于tf-idf权重的实值向量，则会构建一个|V|维实向量空间(|V|是词项数目)。空间中的每一维都对应一个词项，文档被表示为空间中的一个点或一个向量

8.向量空间模型的优点：把文本内容的处理简化为向量空间中的向量，以空间上的相似度表达语义上的相似度；帮助改善了检索结果；部分匹配的文档也可以被检索到；可以基于向量cosine值进行排序并展示给用户

9.向量空间模型的缺点：向量空间模型是一种词袋模型，并假设标记词是相互独立的，但实际上很可能不是。如近义词，同义词可能被认为是不相关的词，与搜索词同义或近义的词不会被检索到，会影响检索结果；向量空间维度非常高且非常稀疏，如果不做其他优化处理检索效率会很差

10.计算向量的方法：空间余弦，而不是欧式距离。因为在向量空间上，查询与文档可能欧式距离很大但内容相近

第九讲 检索排序

1.检索排序的意义：PPT上没有，扯吧

2.精准topK的加速方法：快速计算余弦(不考虑查询词项的权重，不需要对查询向量进行归一化)；堆排序法N中选K(找出具有非零余弦相似度值的文档，从中选出TopK)

3.非精准TopK策略：

索引去除：只考虑查询词项idf值超过一定阈值的文档；只考虑至少包含3/4查询词项的文档

胜者表：对于词典中的每个词项，提前计算出r个最高权重的文档构成胜者表。给定查询q后，对q中所有词项的胜者表求并集生成集合A 并从A中找出TopK

静态得分：为了使排名靠前的文档既是相关的也是权威的，为每个文档赋予一个与查询无关的权威度。最终的文档排名基于相关度和权威度的组合

影响度排序：多个term对应的文档次序不是统一的，在遇到每个词项时得分进行累加计算。有两种思路可以显著降低用于累加得分的文档数目，一是提前结束(扫描了r篇固定数目的文档或者当前记录的tf已经低于某个阈值)，二是词项按照idf降序排列。

簇剪纸方法-预处理：随机选出根号N篇先导者 ，对于其他文档，计算与他最近的先导者并依附于先导者，称之为追随者。查询时，先通过与先导者计算余弦相似度找出r个先导者，再对每个先导者及其追随者进行余弦相似度计算，直到找到TopK。

4.链接分析排序算法：将整个Web看成静态HTML通过超链接而形成的有向图，其中每个网页是一个顶点，每个超链接对应一条有向边。

pageRank思想：vote，被引用多的就是重要的。对每一个网页赋予一个初值，将一个网页的pageRank平摊给所有他引用的网页，构造方程组求得网页级别，迭代近100次后得到近似结果。如果权威页面中有指向其他页面的链接，容易引起主题漂移。

Hits思想：主题相关网页之间的链接对于权重计算的贡献比主题不相关链接价值要更高，专注于改善泛指主题检索的结果，对每个网页都要计算权威值和中心值(Authority页和Hub页)

pageRank与Hits的比较：都是基于链接分析的排序算法，都利用了特征向量作为理论基础和收敛性依据。Hits算法计算的authority只是针对于检索主题的权重，而pageRank的权重是独立于检索主题的。

第十讲 搜索引擎优化

1.SEO：搜索引擎优化，指在了解搜索引擎自然排名机制的基础上，对网站进行内部及外部的调整优化，改进网站在搜索引擎中的关键字自然排名。

2.webspan：通过非法的人为操作误导搜索引擎，使得一些网页的排名高于其自身价值。

具体分为term spam(增加垃圾词项)，link spam(增加垃圾链接)，hiding techniques(超小字体和链接，与背景颜色相同的字体)

第十一讲 信息检索的评价

1.查准率：检索出的相关文档/(检索出的相关文档+检索出的不相关文档)

2.查全率(召回率)：检索出的相关文档/(检索出的相关文档+未检索出的相关文档)

3.计算查全率的pooling方法：对多个检索系统的TopN个结果组成的集合进行人工标注，标注出的相关文档集合可以作为整个文档的集合。

4.精确率：(检索出的相关文档+未检索出的不相关文档)/(检索出的不相关文档+未检索出的相关文档)

为何不用精确率？因为查询相关的文档占很小一部分，所以精确率一般是99%以上，无参考意义

5.F值：召回率R和查准率P的加权调和平均值。 (1+β^2)\*P\*R/(β^2\*P+R)

6.R-查准率：计算检索结果中前R个结果的查准率

7.对多个查询进行评估分为宏平均(Macro)和微平均(Micro)，宏平均是指每个查询的结果求算术平均，微平均是将所有查询视为一个查询综合求一个评估值

8.AP：平均查准率，是指一次查询中求得在每个相关文档位置上的查准率的平均值(一定是相关文档！),分母是相关文档数

9.MAP：多次查询中对于每次查询AP值的宏平均

10.NDCG：normalizing(将检索结果按相关度由高到低排列) discounted(排在前的检索结果的权重比排在后的要大) cumulative gain(将每个文档和查询的相关度进行加和)，标准化折扣累计增益。一种总体观察排序效果的方法，利用检索序列加和的思路来衡量。

11.面向对象的测度方法：

覆盖率：表示检索出的文档中用户已知文档的比例。(查询结果中用户已知的/用户脑海中已知的)，类似于查全率，分母由用户估计

新颖性：表示检索出的文档中用户未知的比例。(查询结果中用户未知的/总的检索结果数量)

多样性：将检索文档按照不同的主题分簇组织，实际输出结果从每簇中均选

第十二讲 相关反馈和查询扩展

1.k-gram：用于英语或拼音的拼写校正。

2.相关反馈：在初始检索结果的基础上，通过用户交互指定哪些文档相关或不相关，然后改进检索的结果。这样可以提高召回率。

3.查询扩展：通过在查询中加入同义或相关的词项来提高检索结果。是一种提高召回率的方法。

4.相关反馈分类：显式相关反馈(用户显式地参加交互过程)，隐式相关反馈(系统跟踪用户的行为来推测返回文档的相关性，从而进行反馈)，伪相关反馈(没有用户参与，系统直接假设返回文档的前k篇是相关的，然后进行反馈)

5.隐式相关反馈的优缺点：不需要用户显式参与，减轻用户负担；用户行为在一定程度上反映了用户的兴趣，因此具有可行性

对行为分析具有较高的要求；准确率不一定能保证；某些情况下需要增加额外设备

6.伪相关反馈的优缺点：不需要考虑用户的因素，处理简单；很多实验也取得了不错的效果

没有通过用户判断，因此准确率难以保证；不是所有的查询都会提高效果

7.相关反馈和查询扩展的比较：相关反馈是局部方法，对用户查询进行局部的即时的分析，查询扩展是全局方法，进行一次性的全局分析来产生同/近义词词典；两者的概念不同；都可以提高召回率；查询扩展可能会显著降低准确率(某一词项有歧义)

第十三讲 隐语义空间

1.SVD：奇异值分解，是线性代数中一种重要的矩阵分解。

特征值分解可以得到特征值和特征向量，特征向量表示这个特征是什么，特征值表示这个特征有多么重要。

矩阵A^T乘以A后得到一个方阵，该方阵的特征值求根就是A的奇异值，AA'的正交单位特征向量构成U，A'A的正交单位特征向量构成V

2.PCA：主成分分析，是一种统计方法，通过正交变换将一组可能存在相关性变量转化为一组线性不相关的变量。转化后的变量叫做主成分。

3.SVD的意义：奇异值分解能够有效地降低数据的维度，还可以保留绝大部分的信息。

可以用于图像压缩，用于解PCA问题，用于推荐算法，用于NLP的算法

4.LSA：隐语义分析，一种信息检索代数模型，使用统计计算的方法对大量的文本集进行分析，从而提取出词与词之间的潜在的语义结构，并用这种潜在的语义结构来表示词和文本，达到消除词之间相关性和简化文本向量实现降维的目的。

5.LSA的意义：将文章和单词都映射到同一个语义空间，可以用于文档聚类和文档分类。发现词与词的关系可以用于同义词、歧义词检测。

6.pLSA：概率潜在语义分析，考虑到word和doc的共现形式，pLSA基于多项式分布和条件分布的混合来建模共现的概率

7.pLSA实现过程：按照概率选择一篇文档d；选定文档后从主题分布中按照概率选择一个隐含的主题类别p(z|d)；选定后从词分布中按照概率p(w|z)选择一个词。总结来说就是选定文档生成主题，选定主题生成词。根据大量已知的文档-词项信息，训练出文档-主题和主题-词项。

8.pLSA的优势：定义了概率模型，而且每个变量及相应的概率分布和条件概率分布有了明确的解释；相比于LSA隐含了高斯分布，pLSA隐含的multi-nomial分布假设更符合文本特性；更好刻画一词多义，用多项式分布描述词频向量；可以利用各种准则来确定topic；pLSA的优化目标是KL-divergence最小，而不是依赖于最小均方误差等准则。

pLSA的缺点：随着doc和term的增加，pLSA模型也线性增加；pLSA无法生成新文档的模型；EM算法需要反复的迭代，需要很大的计算量；概率模型不够完备，不是完整的贝叶斯模型

9.LDA：隐含狄利克雷分布，LDA和pLSA思想上一致，

不再认为各个主题在文档中出现的概率分布和各个词语在某个主题上出现的概率分布是确定的，增加了狄利克雷先验

实现了全贝叶斯化

第十四讲 相似性计算

1.欧氏距离：∑(|xi-yi|^p)^1/p

2.标准化欧氏距离：欧氏距离具有如下缺点，将各个分量的量纲相同看待；未考虑各个分量的分布可能是不同的

标准化欧氏距离是针对欧氏距离的一种改进。x=x-m/s m是均值，s是标准差

3.文本相似度量方法：

String based method：

Character based method：最长公共子序列(LCS)，编辑距离(字符的插入、删除、替换)，扩展的编辑距离(插入、删除、替换、相邻字符的交换)，Jaro方法和Jaro-Winkler方法(考虑两个字符串之间相同字符的顺序位置和个数)，海明距离(适用于长度相同序列之间的比较)

Term based method：余弦相似度，Jaccard相似度，Dice系数(2|X∩Y|/|X|+|Y|)

Corpus based method：LM(Language Model) 通过语言模型来发现近义词、同义词，TM(Topic Model) 通过词与词的共现来反映词与词之间的相似性

4.shingle堆叠算法核心思想：将文件相似性问题转化为集合相似性问题，给定一个正整数k和文档d的一个此项序列，可以定义文档d的k-shingle为d中所有k个连续词项构成的序列，用于进行文档重复检测。

5.LSH：局部敏感哈希，是一种常见的用于处理高维向量的索引方法。指在面对海量数据时，一般的算法无法快速降维查询相似度高的数据子集，利用特定的hash算法，将高维数据映射到低维空间，以较高概率快速寻找相似度高的数据子集。采用过滤—验证框架，在过滤阶段把不可能成果结果的数据对象过滤掉，过滤后的数据对象集合叫做候选集，在验证阶段对候选集合进行相似度计算。

6.MinHash：是LSH的一种，可以用来快速估算两个集合的相似度。每个候选集合都是由一系列词项组成，通过hash函数可以把每个term映射为一个整数，这样我们可以得到一个集合的最小哈希值。minHash又分两种，一种是使用多个hash函数，这样每个集合可以得到k个最小值。集合A、B的相似度可以表示为Ak∩Bk/Ak∪Bk；第二种是使用一个hash函数，这样保留每个集合的r个最小值，依旧是Ar∩Br/Ar∪Br

7.SimHash：是LSH的一种，主要分为五个步骤，分词(将文档进行分词并对每个词项标注权重)，hash(选择合适的hash位数，通过hash函数计算各词项的hash值)，加权(将hash值与权重相乘，若hash位为0则乘-1)，合并(将所有词项的加权结果进行累加)，降维(对于合并后的一个hash值，若某一位大于0则置为1，否则置0，这样作为整篇文档的simHash值)。通过对文档间simHash海明距离的计算，可以判断文档是否相似。

8.词的两种表示方式：one-hot representation，每个词表示为一个很长的向量，向量维度为词表大小，具有语义鸿沟，稀疏，维度灾难，无法表示unseen words的缺点

distributed representation：每个词表示为一种低维实数向量，每一维都可以看作词的语义或者主题信息

9.word2vec：用来产生词向量的相关模型，可以用于将一个词转化为一个词向量。

CBOW：连续词袋模型，根据某个词前面的C个词或者前后的C个连续的词来计算某个词出现的概率。

Skip-Gram：与CBOW相反，根据某个词分别计算其前后出项某几个词的各个概率

word2vec的意义：可以用来查找相关词语的列表；可以用于寻找对应关系；可用于机器翻译；可用于推荐系统

第十五讲 图片检索

1.图像检索的分类：基于图像库的图像检索；基于Web的图像检索；基于文本的图像检索；基于内容的图像检索(CBIR)

2.CBIR的关键技术：图像特征提取和匹配(低层视觉，如颜色、形状、纹理、空间位置关系；语义内容)

3.颜色特征的表示：

颜色直方图：采用一定的量化方法对颜色进行量化，统计每一个量化颜色占整幅图像的比重

优点：具有平移、尺度、旋转不变性；适用于描述难以自动分割的图像

缺点：未包含空间位置信息；不同的图像可能有相同的颜色分布

颜色相关图：用颜色对相对于距离的分布来描述信息

优点：反映了像素对的空间相关性，以及局部像素分布和总体像素分布的相关性

颜色矩：在颜色直方图的基础上计算每个颜色的矩估计，用这些统计量代替颜色的分布来表示颜色特征

优点：不需要颜色空间量化，特征向量维数低

缺点：该方法检索效率比较低

颜色一致性矢量：本质上是一种引入空间信息改进的直方图算法，统计了图像中各颜色最大区域的像素数量

4.感知哈希算法：缩小尺寸，简化色彩，计算灰度平均值，比较像素的灰度，生成hash，进行相似度计算

优点：简单快速，不受图片大小缩放的影响

缺点：图片内容不能变更，如果在图片中加几个文字，可能就认不出来了

pHash：将均值的方法发挥到极致，使用离散余弦变换来降低频率。

5.纹理分析的途径：

基于结构的纹理分析：研究组成纹理的基元和他们的排列规则

基于统计的纹理分析：寻找刻画纹理的数字特征

基于模型的纹理分析：以图像的构造模型为基础，采用模型的参数作为纹理特征

基于信号处理的纹理分析：运用了信号分析的知识来进行纹理分析

6.心理学6个纹理特征：稀疏度、对比度、方向性、线状性、规则性、粗糙度

7.灰度共现矩阵的思想：用邻接像素灰度值的相邻状况来刻画图像的纹理特征，用有限的矩阵关系来表现纹理特征。具体又可以分为水平、垂直、对角线、反对角线四个方向。

8.基于信号处理的纹理特征分析过程：先对纹理图像信号采用频域或者空域滤波处理，再对纹理图像进行分析及解译

9.LBP特征：局部二值模式，结合了纹理图像结构和像素统计关系的纹理特征描述方法。每个像素点都和其相邻的8个像素点做对照，若四周大则为1，若四周小则为0，8个二进制数转换为十进制的LBP编码，用这个值反映该区域的纹理信息。

优点：光照不变性，旋转不变性，灰度不变性。

10.形状的描述符主要分为：

基于轮廓的形状描述符：描述形状目标区域边界轮廓的像素集合

链码(选定一个起始点，按照顺时针方向沿边界顺次找出下一个边界点)，基于网格的方法(将图像形状边界映射到网格的左上角，若某个单元格被形状边界完全或大部分覆盖，则赋值为1，否则赋值0)，距离直方图(在边界均匀取特征点，计算特征点到质心的距离，建立距离直方图)，边界矩(将边界点到质心的距离理解为一个分布，计算分布的矩)，傅里叶描述子

基于区域的形状描述符：描述形状目标区域所有像素的集合

简单的区域标量描述(区域面积，拓扑描述子，投影，离心率，矩形度，细长度)，基于区域的不变矩

11.大津法：证明了类内差异最小等同于类间差异最大。w1 = n1/n；类内差异 = w0\*(u1-u)^2+w1\*(u2-u)^2 = w0w1(u1-u2)^2 = 类间差异

我们要做的就是找到一个T作为区分前景和背景的阈值使得类间差异最大

12.图像局部特征：

HOG特征：方向梯度直方图。

主要思想：局部目标的形状可以被梯度方向密度分布很好的描述。本质是梯度的统计信息，而梯度主要存在于边缘的地方。

主要实现过程：灰度化，gamma校正，计算图像中每个像素的梯度，将图像划分为小cells，统计每个cell的梯度直方图，利用cell的梯度直方图来统计cell的梯度信息(如梯度方向、梯度大小)这样就得到了cell的HOG特征，将几个cell组成一个block，将其中的cell特征串联起来就得到了block的HOG特征，将所有block的HOG特征串联起来就得到了image的HOG特征

SIFT特征：尺度不变特征转换。

主要思想：在不同的尺度空间上查找关键点，通过求一幅图中的特征点及其相关尺度和方向的描述子来得到特征。

主要实现过程：建立尺度空间；在尺度空间中检测极值点，并进行精确定位和筛选；求得每个特征点的位置、尺度和方向；计算特征描述子

13.图像检索局部特征全局化的三种方法：

BOF：Bag Of Feature

主要思想：是一种图像的特征表示方法，类似于bag of words，但这次抽取的不是word而是图像的关键特征feature。

优点：在提取特征时不需要label，是一种弱监督的学习方法。

缺点：完全没有考虑到特征之间的位置关系。

VLAD：vector of locally aggressgated descriptors

主要思想：和BoF类似，分成k个聚合类，只考虑局部特征和它最近的聚类中心

不同的是，不是简单的把这个局部特征归到中心上，而是保留了该特征与最近聚类中心的距离

优点：考虑了特征点每一维的值，对局部特征有了更多的刻画；减少了损失信息

缺点：

FV：fisher vector

优点：减少了一部分损失信息

缺点：局部特征的表示复杂化，增加了计算难度

主要思想：在BoW的基础上，考虑了局部特征到聚类中心的距离，用所有聚类中心的线性组合表示一个局部特征