**CH3网页文字处理**

**问题背景：**文本处理的概念与目的：1.信息检索的基本组件，为后续应用提供支持：(1).将原始文档转化为词项，以建立索引。(2)使面向查询条件的文档匹配成为可能。2.文档处理与查询解析相辅相成。文档处理完整流程：分词(词条化)->去停用词->规范化处理。**词条化处理**：**词条化：**将给定的字符序列拆分成一系列子序列的过程。**词条化的主要任务**是确定正确的词条，并避免标点等因素干扰。 **英文分词的挑战**：词与词组的切分、标点符号的影响(连字符、引号)、专有名词的拆分。**中文分词的挑战**：没有显式分隔符(空格)、中文对虚词的运用、分词歧义、未登录词(人名地名机构名商品名，专业领域术语，新词、变异词，中英皆受影响)、其他类型字符序列(专业术语中文字与符号结合部分C++、新类型字符序列URL、多种语言混杂表达方式awsl)。**三种中文分词歧义**：交集型歧义(AB、BC都是词)、组合型歧义(AB、A、B是词)、真歧义(文字游戏)。**常见分词方法 基于字符匹配的方法**：（词典词条匹配）**正向最大匹配分词(FMM)**：从左至右尽可能查找最长的词，直到当前字符与已经处理的字符串不构成词， 输出已经识别的词，并从识别出来的词后面接着继续查找下一个词（速度快，但错误率高1/169）**反向最大匹配分词(RMM)：**从右至左尽可能查找最长的词，直到当前字符与已经处理的字符串不构成词(错误率1/245) **双向最大匹配分词(BM:FMM+RMM)** ：综合比较FMM与RMM两种方法的切分效果，从而选择正确的切分,有助于识别分词中的交叉歧义。**两种策略**：两种结果直接合并；取分词结果最短那一个。 **基于匹配分词的一般模型**：**ASM(d,a,m)** ：d表示匹配方向，+1为正向，-1为逆向。a:每次匹配失败后增/减字符数，+1为增字，-1为减字 。m:最大/最小匹配表示，+1为最大匹配，-1为最小匹配 。例如，ASM(+，-，+)即正向减字最大匹配(即FMM) 。汉语最大匹配更有用(最小匹配太琐碎)**最少切分分词(最短路径分词) ：**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的问题。 将每个字元视作节点，每个词形成一条边。边权重可都视为1，也可根据词频决定(尽量切出高频词)，结合权重/概率后可视为基于统计的分词方法。拓展方法:N-最短路径法，保留N条最短的路径，以提供更多分词方案。 **基于匹配分词方法的优**：效率高、直观性好；**劣：**依赖词典；维护高质量词典需极大开支；难以应对新生词汇；词汇频率/重要性往往对结果不产生影响。**基于统计的分词方法：**，𝒄是待分词的句子（字串）。而是切分的结果。设为𝒄切分为𝒘的某种估计概率。为𝒄的所有可能的分词方案。所以目的是找到目标词串𝒘，使𝒘满足:,即估计概率最大所对应的词串。**统计分词的一般化过程：**1. 建立统计语言模型2. 对句子进行分词3. 计算概率最大的分词结果。理论上，基于统计的分词方法可以不需要词典，但实际应用中第2步可以采用机械分词方法进行分词，以获得候选的分词集合：既发挥匹配分词切分速度快、效率高的特点, 又利用了无词典分词结合上下文识别生词、自动消除歧义的优点。**N-gram**指一个由N个单词组成的集合，各单词具有先后顺序。 **N-gram模型的马尔可夫假设：**当前状态出现的概率仅同过去有限的历史状态有关，而与其他状态无关。文本中第N个词出现的概率仅仅依赖于它前面的N-1个词，而与其他词无关。**常见的N-gram模型：**N = 1，一元文法模型（最大概率模型）：. , P(wi)=wi在语料库出现次数n/语料库总词数N。N = 2，Bigram模型： 。N = 3，Trigram模型： **N-gram模型的分词过程:以Bigram模型为例**：1.构造训练语料库，计算所有的与。2.对于每一个可能的分词序列，计算以下公式 :其中，3. 返回最大的所对应的分词序列作为结果。**基于统计文法模型的优：**减轻了对于词典的依赖性(但并未消除，取决于性能与效率的平衡：深度结合匹配分词，效率升但依赖词典；减少对词典依赖，则需遍历更多组合，解空间巨大)**劣：**依赖已有数据中词频的统计，对于新生词汇或专业词汇不友好：冷门领域稀有词汇难以准确划分；易受数据集先验偏差影响。**基于序列标注的分词方法：**基于统计模型的分词方法，进一步抽象可得到一个序列标注问题。四类标注:B(词的开始)、M(词的中间)、E(词的结束)、S(单字词) 。**隐马尔科夫模型(HMM)** **基本思想**：根据观测值序列(字符序列)，找到真正的隐藏状态值序列(标签BMES序列)。**五要素：**2集合（观测集、隐藏状态集），3矩阵（初始状态概率矩阵：第一个字属于某种隐含状态(BMES)的概率，隐含状态转移概率矩阵：各种隐含状态(标签)之间的转移概率，观测状态概率矩阵：从隐含状态到观测值的转移概率）**问题形式：**观测值w1…wn,希望找到标签序列s1..sn,使得P(s1..sn|w1..wn)概率最大。**基本假设：**齐次假设（当前隐藏状态只与上一个隐藏状态有关），观测独立性假设（观测值间相互独立，只与对应隐藏状态有关）**求解方式：维特比算法**(初始化、递归、终止、回溯：典型动态规划)。**条件随机场模型(CRF):** 隐马尔可夫模型的独立性假设难以描述字词之间的复杂关联。CRF具有表达长距离依赖性和交叠性特征的能力、所有特征可以进行全局归一化，能够求得全局的最优解。**长短时记忆模型(LSTM)：**传统神经网络中：每层内部的节点间无连接，所以无法利用上下文关系。循环神经网络（RNN）可利用上下文关系，但当需要较长上下文关系时，对信息依赖的学习能力有限。LSTM通过四层神经网络代替RNN单一神经网络层，使其拥有增加或减少信息的能力。**LSTM+CRF：**传统LSTM+SoftMax忽略了预测序列的标签之间的关联性，可能导致错误标签列的出现。结合LSTM和CRF将有效利用句子级别的标记信息,输出的将不再是相互独立的标签，而是最佳且合理的标签序列。**分词可能带来的隐患：**分词带来的大量低频词，导致严重的数据稀疏；越来越多OOV(Out of Vocabulary)词；分词中难免的错漏将导致额外的噪声；深度学习发展，分词的收益愈发有限。**停用词处理：**停用词，指文档中频繁出现的数字、副词等与语义关系不大的词。**为什么要去除停用词：**重复率很高，会造成索引中的倒排表很长，影响查询性能；对最后结果的排序没什么贡献，反而可能产生干扰。 **停用词类型：** 停用词的设置与语料库的性质有关(通用停用词表,专用停用词表(特定学科/领域)**。停用词识别方法:**文本频率、词频统计、熵计算;结合统计与句法或内容分析。**去除停用词可能导致的隐患**：有些在特定场景下有意义(非、稍微) ；有些停用词的组合是有意义的(的士)，依赖于分词效果。**未来停用词的使用趋势：**逐渐减少使用。现代搜索引擎更关注利用语言的统计特性来处理常见词问题（压缩技术：降低停用词表的存储开支；引入词项权重，将高频词的影响降至最低；索引去除技术，将低于权重的词项排除）**规范化处理：归一化处理** ：**文本规范化的意义**：文本处理的主要目标在于**优化查询词与索引词之间的匹配**，但两方文本内容都可能有干扰(见要考虑的问题)，**规范化的目的**在于尽量保证索引词项符合用户查询输入。**规范化要考虑的问题:**大小写、标点、缩写等规则化处理；词根化处理：词干提取与词形还原；拼写错误检查与修正；同义词/相关词识别与处理；其他类型的文本规范化问题：特殊文本形式(9月19日与9/19),跨语言问题(中文与日语字同)。**词根化处理：**指还原词语的特殊形式的过程，可有效降低词项的数量并减少歧义。**词干提取**：去掉单词前后缀，获得词根（复数、过去分词、进行时）；**词形还原：**单词转变成最基本形态(基于词典，is->be)。**词干提取与词性还原相同点**：1. 目标一致(归并为基础形式)2. 结果交叠3. 方法类似：主流方法均是利用语言中的规则或词典实现 **不同点**：1. 原理上：词干提取：缩减； 词形还原：转变2. 复杂性上：词形还原还需考虑词缀转化、词性识别等3. 实现上：词干提取：利用规则变化；词形还原：更依赖词典4. 结果上：词干提取：不一定得到完整单词；词形还原：完整单词。**错误拼写检查：**采用基于词典或编辑距离的方式进行检查和校对。**编辑距离：**指两个字符串之间转换所需最少编辑操作步数。**一步编辑操作**：替换、插入、删除一个字符。**同义词/相关词处理**：比词根化和拼写错误检查更难处理，借助人工维护的知识库来获取各种词项间的关系(同/反义词，is-a,is-part-of)。**相关应用：**基于同义词/相关词，拓展查询条件

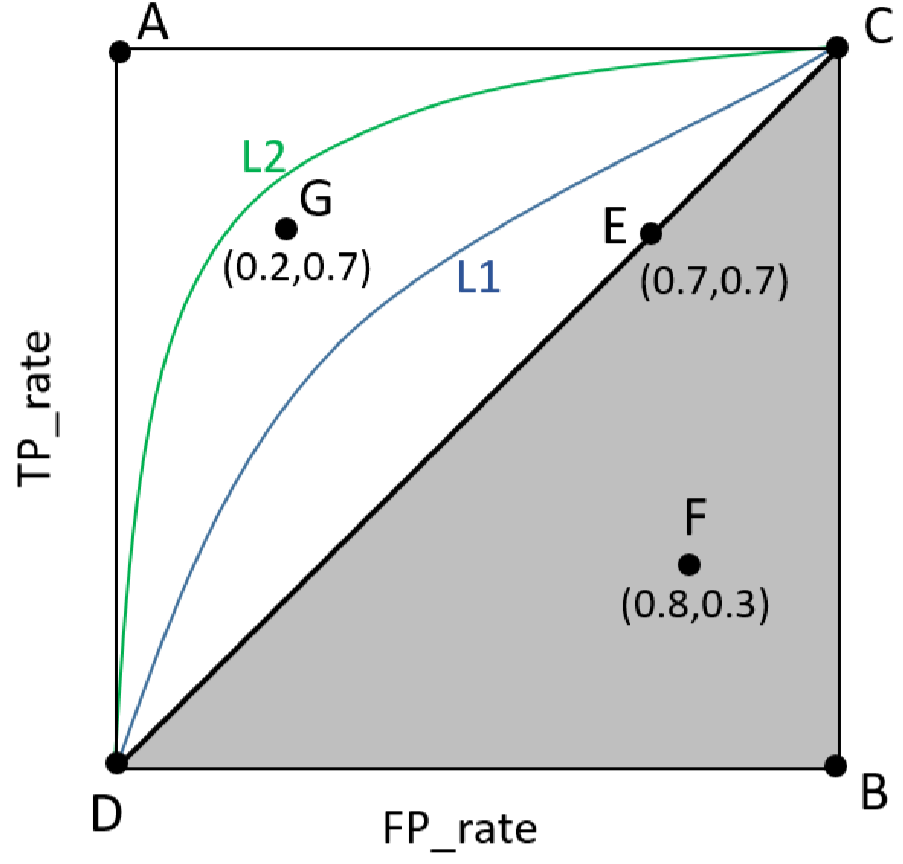
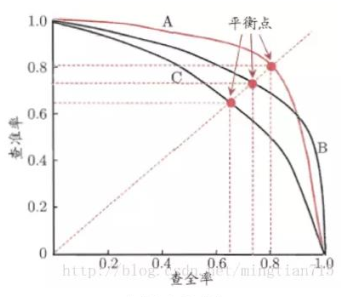
**Ch4网页索引**

**问题背景：网页索引的目的：**索引的质量，关系着整个搜索引擎系统的精度与效率（查询速度、存储空间）关键词查询-》索引与匹配-》返回匹配文档。**索引词项的选择范围：人工索引：**质量好，查询有针对性，更准确反应文档核心内容，但不够全面，依赖人力(大规模文档不行) 。**自动索引：**依赖算法自动分析文档(部分索引：如标题、摘要、关键词等;全文索引：对文档中所有词都进行索引(普遍用)),算法提取索引词，高效全面，但存在误导,或限于算法效果而出错。**索引词项选择原则**：理想状态：表达文档核心内容的语义单位；依赖文本处理技术进行处理，如分词、规范化等。**布尔检索：**文档被表示为关键词的集合。所有的查询式都被表示为关键词的布尔组合。 采用“与、或、非”关系加以连接。相关度计算 ：一个文档当且仅当它能够满足布尔查询式时， 才会将其检索出来。 检索策略是二值匹配（完全匹配）。 **优点：**查询简单，易于理解;布尔表达式可以方便地控制查询结果;可通过扩展来包含更多功能(如文档类型、文档领域、词频、位置信息等)。**缺点：**功能较弱，不支持部分匹配;所有匹配文档均返回，不考虑权重和排序;很难进行自动的相关性反馈 (如根据用户点击量来进行排序)。**实例：1.** OR：可以利用同义词表自动扩展。**2**.m AND (q OR p), 笨办法即从头到尾扫描所有剧本(缺陷：速度慢(尤其对大型文档)；NOT操作的处理需遍历全文；不支持检索结果的排序) **3.关联矩阵**(事先给文档建立索引)：m\*n矩阵，m表示所有文档共出现m种单词，n表示共n篇文档。每一列都是0/1向量，代表该文档里包含的单词，1代表包含这个词，0代表不包含。将给定的查询条件转化为行向量运算：取出查询式里3个单词对应的行，根据运算符求与或非,结果里为1的位置对应结果文档。**关联矩阵缺陷：**稀疏（可以选择只记录1）。**倒排索引：** 包括：**词汇表(词典)**:词项的集合、**倒排表(Posting List):**文档ID列表，列举词项在哪些文档中出现.**倒排索引VS正排索引：**正排：关键索引为文档，易维护，搜索耗时长；倒排：关键索引为词项，构建维护成本高，搜索快。(索引有一定规模时，可压低其更新频率来减少维护开销，另一方面，搜索频率远高于文档更新频率，所以倒排更优) **建立倒排表：**1. （词条化等处理后）检索每篇文档，获得<词项，文档ID>对，并写入临时索引。2. 对临时索引中的词项进行排序(所有文档拉通排)。3. 遍历临时索引，对于相同词项的文档ID进行合并。**查询实例**：A AND B，表长分别为x和y，复杂度O(x+y)，**索引算法**：从两表第一项id开始比较，小的往后走，若相同则查询成功，加入结果集后双方都往后走。**动态索引问题：**文档的添加、删除、修改。考虑使用**主从索引**：维护一个大的主索引，新文档信息存储在一个小的辅助索引中，检索时同时遍历两个索引，并进行合并，如需删除操作，可利用一个新的无效位向量，用其过滤结果（比如对删除的词也用一个倒排表记录该词在哪个文档中的记录已无效）,定期将辅助索引合并到主索引中。但**存在的问题：**1. 频繁合并导致大开销2. 合并过程效率很低(考虑每个词项的倒排表单独形成一个文件，这时合并较为简单，但此时文件读写开销大（压力转到O/S上）)。现实里折中方案：对大索引记录表切分，小的索引记录表合并。**倒排索引优化：处理查询的最佳顺序**：（含OR和AND运算,**not>and>or**）1. 获得所有词项的文档频率（即在多少个文档里出现）。2. 估计出每个OR操作后的结果大小，考虑最坏情况x+y。3. 按照结果**从小到大**（文档频率小的先AND）的顺序执行AND 。**跳表指针：算法：**也是从两表第一个id开始比较，相同时加入结果集，双方后移一位；不相同，表A对应id小，如果该位置有跳表指针，且跳过去后的id仍<=表B的对应位置，那就跳，否则只往后移一个。**设置策略：**设置较多的指针+较短的步长→更多跳跃机会，耗费更多的存储空间。设置较少的指针+较长的步长→更少的指针比较次数 ，存储空间消耗少，但跳跃次数也少。**启发式策略:**如果倒排表长度为L，则间隔sqrt(L)均匀放置跳表指针。若还要**考虑查询词项的分布**：不同的id区间项数密度不同；**索引的动态变化**(文档删减)也要影响。**倒排表的扩展性问题：**1.倒排表中加入更多元素(词项频率、词项类型)2.更多需求：词组查询、词项邻近、文档区域(作者是A的文档)。**短语查询解决方案**：**1.二元词索引**(但词汇表会迅速增长)、多元词索引或划分成多个短查询处理(ABCD变成 AB AND BC AND CD)。**2.位置信息索引**：在倒排表中加入词在文档里出现的位置(文档1：位置1，2，3)，短语查询时用AND处理，同时考察位置情况。代价：存储太大。好处：能知道短语是否出现并确定位置；能用于邻近搜索(如要找的两个词间间隔k个词) **索引存储：**顺序存储 、Hash Table 、B+-Tree、Trie树 (前缀树)。**索引压缩：（从词典，倒排表两个维度）**意义：节省磁盘空间、提高效率(内存中快)。**Zipf 定律**:排名第 i 多的词项的文档集频率与 1/i 成正比 。**词典存储：定长存储**(词项定长20B，文档频率和倒排表指针各4Byte)。**压缩词项列表：**将词典视作单一字符串：词项之间用指针分割。**按块存储(进一步压缩)**：通过为每k个词项存储一个指针,来减少指针的总数量. 每个词需要额外1个字节用于表示词项长度(5smile3dog2is).每一组可节省k-1个指针开销(每个约3字节)，损失k字节存储词项长度。**K值的选取**：按块存储前是标准二分法；按块存储后，二分法只能在块外进行，块内是线性查找，k太大，线性查找部分增多，效率更低。**前端编码：**利用连续词项间的公共前缀。**倒排表存储**:**ID间距代替文档ID**：数值更小。**可变长度编码**：1. 采用间距存储。2. 对于一个间距G，先存储G，并分配1bit延续位。3. 如果G<128，则采用第一位延续位为1 + 7位有效二进制编码的格式。如果G>=128，把二进制码7bit切分，最高一段不足的补0，最后一个字节(8bit)的延续位为1,其他字节延续位为0.**优点：**相比于4字节整数，可变字节码在小数字上的短码可以节省更多空间

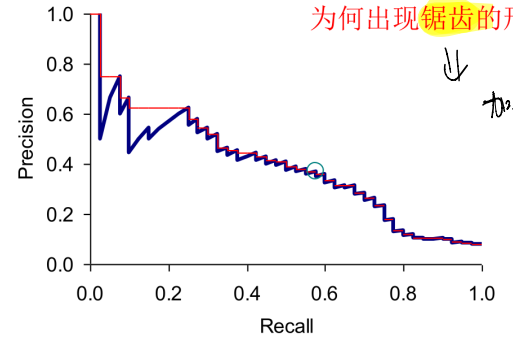
**Ch5查询**

**更一般的检索方式：**文档与查询的相关性排序返回合适的文档(eg:TopN)。**查询表达理解：查询条件难以理解的原因：**用户表达的精简性和歧义性(缺乏精准描述)、用户表述方式的差异性（方言、同义词、错误）。**理解用户查询的方式**：1.最基本的途径:基于查询的自然语言理解。2.引入相关性反馈：用户直接对查询结果进行评价。3.引导用户表达真实查询意图(查询建议与查询扩展) 。4.借助其他信息，完善对于用户的理解(用户间接性反馈、情境信息与情境建模)**相关性反馈：**用户在查询后标记相关/不相关，然后迭代更新查询，以获得更好的结果。**基本流程(下述步骤会多次迭代趋于最优)**：1. 用户提出一个查询条件。2. 对于返回的文档，用户标出相关与不相关的部分。3. 系统根据用户反馈，获得用户信息需求更为准确的描述。a) 基于相关性信息，更新查询条件，如为不同词项添加不同权重，或在查询条件中添加新的词项。b) 基于新查询条件，获取新的结果文档并再次提交用户进行评估。**常用技术：衡量查询与文档间的相关性：向量空间模型(VSM)**：查询和文档表征为词项权重构成的向量->通过向量间的相似度评估相关性。**一类文档的质心：**所有文档向量求和除以文档数。**基于向量模型的罗基奥(Rocchio)算法：目的：**使得查询向量尽可能离相关文档更近，离不相关文档更远。但难以获得完整的相关/不相关文档集合。为已知相关文档的向量集合，为已知不相关文档的向量集合。为初始查询向量. 为权重，根据手工调节或经验设定。**正负反馈：**正反馈的价值往往大于负反馈 （用户更关心符合需求的标准答案）。可以通过设置 *beta > γ* 来给予正反馈更大的权重。甚至只允许正反馈，即 *γ = 0* 。收集真正的负反馈往往比较困难。**相关性反馈中的基本假设：假设A1:**对于某个查询，用户知道在文档集中应使用哪些词项来表达。用户必须用足够的知识来建立一个不错的初始查询。不成立的情况:用户不知道如何表达，或词汇表与文档集词汇表不一致。**假设A2:**相关文档中出现的词项类似，因此可以通过相似文档来相互搜寻。在高维空间中，相似文档应该能够形成一个紧密的簇。不成立的情况:文档词汇表不一致，或是通用概念的不同实体。**相关性反馈存在的问题：1.**相关性反馈可能影响用户体验：用户不愿意提供显式的相关反馈；用户不希望因为相关性反馈(迭代)而显著延长搜索时间 。**2.**相关反馈生成的新查询往往很长，降低系统效率，增加计算开支：一种做法是只改变重要词项权重而不增加新词项，但效果有限。**3.**有时经过相关性反馈后，会返回某篇特定文档: 被相关性反馈捕捉的词项，未必是用户需要的内容。**相关性反馈分类：**显式反馈(用户显式地参与交互过程)。隐式反馈(系统追踪用户行为来推测返回文档的相关性)。伪反馈(在没有用户参与的前提下，直接假设返回结果是相关的，并进行反馈)。**查询扩展：**用户针对词项的合适程度给出反馈，这些反馈将被用来构建更为完整的查询条件(可将具有歧义的查询词展示出来：输入周杰伦，会冒出来新歌首发等选项)。 **类型：1.拼写错误处理**(基于词典或编辑距离来检查校对)**2.同义词扩展**：有助于提高查询的召回率。词典维护编纂代价大。**查询扩展的类型：**利用人工编纂的同义词词典、全局分析与同义词词典的自动生成(基于统计词汇间共现关系)、基于搜索日志进行优化。**自动构建同义词典思路：**如果两个词经常和相似的词共同出现，则它们很可能是相似的、如果两个词经常与相同的词以一种特定的语义关系共同出现，那么他们很可能是相似的。**基于共现的同义词典构建：**给定词项-文档矩阵A，计算C=AA^T，对A需进行归一化，使行向量的大小为1，对于每个词项，选择C中相似度最高的若干词项作为同义词。**情景感知的查询理解：**用户查询存在歧义、过精简等问题，需要借助其他信息(查询上下文(历史等)，查询时环境信息)来协助判断。**情境信息**：用于提供更精确的信息检索和过滤服务。**查询上下文：**能更好理解用户查询词，但如何拆分查询会话(如连续多个搜索记录哪些才是相关的)。**基于上下文感知的搜索的基本流程：线下训练阶段(模型准备)：**查询词归纳为查询概念(避免查询词稀疏性影响)；建立模型描述查询概念间的关联关系。**线上服务阶段（感知查询阶段）：**切分会话，判断与当前查询相关的上下文查询与点击记录；根据已有查询记录理解用户当前意图，进行精确查询。**查询概念：**一组有着相同语义的查询词(可解决查询词稀疏性问题)；启发式方法:Query-URL的二部图聚类(多个查询URL归纳为一个Query)。**基础的上下文感知方法：**将会话内查询序列转化为查询概念序列，考虑特定长度以内的所有序列模式，保留频率高于阈值的模式，并存储为后缀树，当面临上下文感知任务时，根据已有序列找到相应节点，从而获得候选查询建议.**进阶的上下文感知方法：**基础方法虽然有效利用了上下文信息，但将查询词限制在了一个查询意图里，同时仅能推荐查询扩展，不能帮助判断文档相关性。引入可变长度HMM，用户意图(隐变量)，查询上下文和点击记录为观察值。**上下文感知的不同类型：1.查询重组:**用户后续查询只是先前查询的重新表述，目的不变或类似(为了得到更多结果；先前点击的内容不会再点击)。**2.查询特化**：对先前查询的部分内容作更具体深入的查询(泛指内容略去)**3.查询泛化：**对先前查询的部分内容进行更泛化的查询，即不局限于特定话题，而是对该查询更广泛的兴趣(游戏网站->游戏介绍和历史)。**4.一般关联**：根据先前查询，补全用户在当前搜索中的特定意图(即先后的查询词上没有重叠，可能上下文相关/不相关，如先搜游戏机再搜FIFA，大概率FIFA游戏)。**更复杂的情境信息：多元信息结合解读。例子：借助情景感知的模式化行为减少开支**

**CH6排序**

******信息检索与相关度计算：信息检索模型概述：**是用来描述文档与查询的表示形式与相关性的框架，信息检索的实质是对文档基于相关性进行排序。**模型的形式化表述**：**[D,Q,F,R(Di,q)]**，D文档表达，Q查询表达，F查询与文档间的匹配框架，R：查询与文档间的相关性度量函数(Di特定文档，q特定查询)。**信息检索模型分类：**1.基于集合论的模型(布尔模型)2.基于代数论的模型(向量空间模型)3.基于概率论的模型(概率/语言/推断网络模型)。 **相关性可通过以下两个维度进行衡量**：主题相关:文档与查询在主题上的一致性。用户相关:文档在多大程度上满足用户需求。**布尔模型**：D:词项组合(不是单词)；Q布尔表达式(词项+布尔运算符)；F完全匹配(二值匹配)，R满足布尔表达式为1，否则0。**从完全匹配到部分匹配的改变：量化方法:Jaccard系数**：,计算两个集合重合度。缺点：仍将文档视作词项集合，查询词本身未做重要性区分，不考虑词项频率，没有仔细考虑文档的长度因素，罕见词比高频词的信息量大但此处未考虑。**词项频率TF**：词项频率，指词项 t 在文档 d 中出现的次数。对数词频： ,缓和了数量级的差异性造成的影响，可以把文档与查询匹配得分定义为**所有同时出现在查询和文档里的词项的对数词频之和**。(考虑在关联矩阵的基础上把0/1值换成词项频率)。考虑到停用词这种出现得多但没啥用，所以要找有区分度的**罕见词**。**文档频率DF**：，指出现词项 t 的文档数量。逆文档频率idf：（N是文档总数），罕见词df\_t小，但是idf\_t高,过于频繁的词(the)作用被抹掉。**tf-idf：**以上两者相乘，**在少数文档里多次出现的词**来衡量文档相关性。同样可用于关联模型**向量空间模型(VSM)：**每个文档和查询视作一个词项权重构成的向量。查询时通过比较向量之间相似性来进行匹配。**形式化描述：**D文档/Q查询 均表示为向量，每一维对应词项的tf-idf值(查询Q一般用one-hot)，F:非完全匹配，R：两向量间相似度度量。**相似度计算：**欧式距离(对向量长度非常敏感，或者说数据的均值/方差产生了影响，趋势/偏好相同的反而距离更大；考虑作了归一化再用)；**余弦相似度**：cos(q,d)=夹角越小->值越高，相似度越高。**向量空间模型的流程：**1. 将文档与查询表示成词项的tf-idf权重向量。2. 计算两个向量之间的相似度。3. 按相似度大小进行排序，将Top-K的文档返回给用户。优点:简洁直观，可以支持多种不同度量或权重方式，实用效果不错。缺点:缺乏语义层面的理解和匹配，同时依赖tf-idf值也可能造成干扰：用户无法描述词项之间的关系，词项之间的独立性假设实际上不成立。 **概率模型：**按照文档与查询的相关性概率大小来排序，描述方式:P(relevant | document, query)，词项出现在相关文档的概率依赖统计，也可通过语义的概率表达。**网页权威性计算：**

**PageRank思想：**将网页(或文档)视作一个点，网页之间的超链接视作一条边，形成有向图。点之间的有向连接表示网页之间相互引用，相互推荐的关系。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。优质网页引用或推荐的网页，一定还是优质网页：**三重衡量标准：**链入链接数：单纯意义上的受欢迎程度；链入链接本身是否推荐度高：欢迎本身是否具有权威性；链入链接源页面的链接数：被选中的几率，体现源网页是否滥发推荐。 。为网页的PageRank值。为指向网页的某个网页的PageRank值。为网页发出的链接数量。d为阻尼系数，取值在0-1之间。N为网页总数，为链入的页面集合。**PageRank的迭代计算过程（实际是个马尔科夫过程）：**1. 给每个网页赋予一个初值1/N。2. 利用之前的公式进行迭代有限次计算，得到近似结果，若某轮迭代前后有某一维的PageRank差值高于阈值，则继续迭代，直至收敛。**简单计算方式：**从邻接矩阵A到跳转矩阵M：倒置后，将各个数值除以所在列的非零元素数。然后利用初始PageRank值P（列向量）不断进行P\_n+1=M\*P\_n的迭代。**特殊情况：**陷阱节点(只有一条指向自己的边即环，无出边)、终止节点(没有任何出边)、 孤立节点(没有任何入边，仅有初始概率，不在更新，也不影响其他结点)。**Restart机制：**其中的(1-d)/N的部分，相当于以一定概率重新选择起点 。此时所有节点以等概率被选中->跳出了陷阱和黑洞的干扰（d一般取0.85）。**PageRank收敛性：A=dM+[(1-d)/N]\*E, A为随机转移概率矩阵，1-d为随机跳转概率。**M为跳转矩阵，E为全1矩阵(与M同维度)，其余同前面公式。所以迭代过程变为**P\_n+1 = A\*P\_n**。马尔科夫过程的三个收敛条件：转移矩阵M为马尔科夫矩阵(M矩阵所有元素都>=0，且每一列元素和都为1)>转移矩阵M为不可约的(当图是强连通时，M为不可约，而Restart保障了这一条件) 。转移矩阵M为非周期的（某一状态经若干步跳转后又回到该矩阵，即A^n=I,则为周期的）。这三个条件，PageRank算法都满足，保障了收敛性。 个性化PageRank：用户视作网络中的一个虚拟节点。任何跳转行为均从用户开始，即用户自身作为浏览行为的起点。从用户出发的跳转概率不一定视作均等，可根据用户对不同网页的偏好决定其概率。基本公式形如 ***r= (1*一*d)Mr + dv*** 的形式，v 体现了用户偏好，可视作不同网站在用户偏好下的点击概率。但代价过大。主题敏感PageRank：考虑偏好因素的同时减轻计算资源负担。每个网页隶属于某个特定主题。在计算部分，区别主要在于起始节点和Restart部分。对于某个Topic来说，起点仅限于隶属于该Topic的网站。(1-d)e/N更改为了(1-d)s/|s|，s为一个N维向量，如果某个网站隶属于该主题，则 s 中的该维为1，反之为0 ，|s| 表示 s 中为 1 的元素的个数。最终，每个网站得到一个向量，代表每个topic下的Pagerank值，即使不属于某个Topic，网页也可以通过PageRank获得相应的数值。同样，用户的需求也将表示为Topic向量。两个向量的内积可以反应这一网站在这种主题需求下的权威性。 **Hilltop算法：**过于依赖PageRank，会鼓励增加大量无效链接而提高PK值。**从属组织网页：**182.61.200.X的IP地址或者XXX.ustc.edu.cn。**专家页面:**与某个主题相关的高质量页面.专家页面一般通过如下启发式规则进行挑选：至少包括K个出边， K由人为指定；K个出边指向的所有页面相互之间的关系都符合“非从属组织页面”。专家页面之外的部分被称作**“目标页面集合”**,即需要进行PageRank计算,判定其权威性的部分。**算法基本流程:**根据用户查询需求，选定相关的专家页面，并计算其相关性。遵循PageRank算法，对目标页面集合进行排序，即将专家页面的相关性得分，传递到其他的目标页面上去。整合最相关的专家页面和得分较高的目标页面返回给用户。PageRank总结：PageRank是一种基于链接分析的全局网页排序算法。优点：对于网页给出全局的重要性排序，并且可以离线完成以保障效率 。基础的PageRank算法本身独立于检索主题，可以实现结果的通用。缺点：主题无关性，同时对于恶意链接、广告链接等缺乏区分。旧网页得分更高，因为新网页往往少有入链。一般不能单独用于排序，需要与相关性排序方法相结合。效率较低。**HITS算法**：引入了网页角色区分.**权威网页**(Authority指某个领域或某个话题相关的高质量网页),**中心网页**(Hub指向了很多高质量的权威网页)。算法目的即找到与查询主题相关的这两类网页，尤其是权威网页.**基本假设:**1. 好的Authority会被很多好的Hub指向。2. 好的Hub会指向很多个好的Authority。,共两个权威值需要计算.**HITS的计算过程：**1. 根据关键字获取与查询最相关的少数页面，及与这些页面有链接关系的页面,作为待选集合。2. 对所有网页的 a(p) 与 h(p) 进行初始化,可都设为1 。3. 迭代计算两个步骤，直到最终收敛。4. 输出Authority或Hub值较高的页面。(M为邻接矩阵)注：每**次迭代同时求出a和h后需要对两者分别进行欧式归一化(即向量求平方和来做分母)，保持和为一。** HITS总结：HITS算法是一种能够区分网页功能的排序算法。优点：更好地描述互联网组合特点；主题相关，因此可以单独用于网页排序。缺点：需要在线计算，时间代价较大。对链接结构变化敏感，且依然可能受到链接作弊的影响。**排序学习问题：网页内容的排序问题**，本质上是**学习排序器**的问题:输入网页（文档）内容及查询，输出该网页合适的排序位置。排序学习是个有监督学习问题，基于已知的排序（如用户反馈）,为新网页-查询对给出排序。**查询决定了排序学习的逻辑结构**:排序的损失函数应设计在查询层面， Ranker与Query对应;位置更为敏感，排在前部（尤其顶端）的部分更重要(用户只看开头页面)。评价指标基于排序结果进行，不是简单的分类问题。**分类**：**Pointwise**，将排序退化为分类或回归问题，**输出:**网页对应的分类(有序)、回归值或有序回归值。**评价：**Pointwise类排序算法可以简单且广泛地套用已有回归、分类算法，但其局限性也较为明显:1. Pointwise类方法往往更为注重文档的相关度得分，而并不注重文档之间的相关性排序,Pairwise类方法的出现，为解决这一问题提供了新的手段其次 2.不同查询所对应的文档，尤其相关文档数量不同，对损失函数的贡献也各不相同，一定程度上影响效果。**Pairwise:**比较一对网页之间的相关度(两两比较，相关的往前靠，是二分类或三分类问题)，比如相关 > 不相关，**输出:**网页对之间的偏序关系。可用Ranking SVM。**评价**：相比于Pointwise类算法， Pairwise类排序算法通过衡量样本之间的偏序关系，实现了从绝对相关性（分值）到相对偏序的改进。**缺陷：**1.两两成对导致样本数大为提升，计算资源开支增加2.Pairwise类算法仍然受样本不平衡问题的影响(如果不同Query的Doc数量相差很大，从偏序的对来说， Doc（对）数量较多的Query将掩盖其他的Query,从而产生干扰)3.Pairwise类算法无法体现全局排序的合理性(某个成对文档排序错误，发生在不同位置，在Pairwise类条件下是等价的,eg:7篇文档，相关性从1到3，理想状态下的排序为3 2 2 1 1 1 1,假如有如下两个排序：Rank A： 2 3 2 1 1 1 1 Rank B： 3 2 1 2 1 1 1这两个排序，在Pairwise类算法中是等价的但后者显然更合理**Listwise:**对整个网页集合进行排序 ，**输出:**整个集合的完整排序，往往依赖特定排序指标。方式：1.采用某种IR指标对排序件进行优化(此类方法较为直观，且可以直接优化所获得排序结果的衡量指标,但面临明显的挑战：大多数排序指标，如NDCG等，因为与排序相关，属于非光滑、不可微的函数。所以传统的优化方法很难直接应用于该问题)，可以考虑AdaBoost的变形AdaRank，利用弱排序器组合实现全局排序2.直接设计面向完整排序的损失函数。**缺陷：**两个网页并列的情况如何处理。

**Ch7结果评价常见内容：性能**(返回多少相关文档，是否遗漏，排序是否靠前)、**效率**(响应速度，时间空间开销)、多样性、权威性、时新性、更新频率。**评价前提**：文档集合，查询 主题集合，评价指标相同，对不同检索系统/排序方法进行比较。**单查询评价:无序结果评价:**TP：样本为正例，且被判定为正，即真正，检索出的相关文档**。**FN：样本为正例，但错误地被判定为负，即假负，未检索出的相关文档**。**FP：样本为负例，但错误地被判定为正，即假正，检索出的不相关文档**。**TN：样本为负例，且被判定为负，即真负，未检索出的不相关文档。(P/N:表示算法对样本的判断；T/F:表示算法判断正确与否)**准确率(Precision)**：检索出的文档中，相关文档所占的比例，也称**查准率**。TP/(TP+FP) 。**召回率(Recall)：**所有相关文档中，被检索出来的部分的比例，也称**查全率**。TP/(TP+FN) 。Accuracy：(TP+TN)/(TP+TN+FP+FN),模式分类里常用，但信息检索任务不常见，因为人们只关注相关的，负样本无意义;而且信息检索里不相关文档占比极大，接近1(eg:搜索引擎对于查询直接不返回任何结果，此时Accuracy=TN/(TN+FN))。应用场景：**分类问题一般都用准确率召回率衡量。**邮件分类（较高准确率）,智慧医疗（较高召回率）召回率近似计算：缓冲池方法（针对某一检索问题，各个算法分别给出检索结果中的Top N个文档(N常取50-200)，将这些结果汇集起来并进行人工标注，从而得到一个相关的文档池.潜在假设：大多数相关文档都在这个文档池中，可用于比较各个算法的相对优劣 ）**.F值：**准确率与召回率的加权调和平均数： (即取α=0.5或β=1,两者同等重要)，**不用算术平均的原因：**调和平均较为保守，在结果上小于算术平均或几何平均，算术平均和几何平均在处理极端情况下效果不够合理(返回全部文档的搜索引擎F>=0.5,当用算术平均时)。P-R曲线：以准确率和召回率分别作为两条轴线，通过选定不同的阈值得到不同的P-R点并连接成线。通过P-R曲线，可直观地看出准确率 与召回率之间的平衡关系。仅选取了一定量的点来作线.**ROC曲线(接受者操作特征曲线)：**以真正率(命中率) TP / (TP+FN)和假正率(误报率)FP/(FP+TN)作为两条轴线，通过选定不同的阈值得到不同的真正率- 假正率点并连接成线。对角线表示区分能力为0，即随机猜测。在对角线上端越远，效果越好。低于对角线的结果无意义(无区分度) 。**基于曲线判别好坏：**曲线下**面积（ AUC）**更高者效果更好。当使用P-R曲线时,可使用平衡点（Precision = Recall）计算，值越高越好。两种曲线的选择：ROC曲线兼顾正负样例，更为全面，而P-R曲线则只考虑正例。用户更关心正样本，如果面向特定数据集，P-R曲线是个好选择 。P-R曲线受分布影响大，若有多份数据且正负比例不一，ROC曲线更合适。正负样本比例失调时(且仅一份数据)，P-R曲线更合适。当负样本比重过高时，ROC曲线难以体现出性能的差异性.**有序结果评价:**P@N(Precision@N)指前N个检索结果文档的准确率(相关数/N)，如果相关文档数小于N，P@N的理论上限可能小于 1. R@N (Recall@N)指前N个检索结果找回的相关文档占总相关文档的比例，理论上限为N/相关文档数。R-Precision：检索结果中，在所有相关文档总数位置上的正确率，eg:相关文档总数为3，则考察P@3作为R-Precision。**面向P,R@N的P-R曲线：**在有序结果情况下，可以不再采用不同阈值作为P-R值的依据，而是通过依次计算前N个结果对应的P-R值绘制曲线。**出现锯齿状**是因为新的不相关文档被检索时，Recall不变，Precision下降(新加的是相关文档时，P,R均提高)。**平均准确率(AP):**对不同召回率点上的正确率进行平均。**未插值AP**(某个查询Q共有6个相关结果，排序返回了5篇相关文档，其位置分别是第1,2,5,10,20位,则**AP=(1/1+2/2+3/5+4/10+5/20+0)/6**).**插值AP**：事先选定插值点数并进行插值(在召回率分别为0,0.1,0.2,…,1.0的十一个点上的正确率(P@N,召回率为0时P@N=1)求平均,等价11点平均).**简化AP**:只对返回的相关文档进行计算， AP= (1/1+2/2+3/5+4 /10+5/20)/5，倾向那些快速返回结果的系统， 没有考虑召回率和补零的情况.**相关度分级：前面的指标都基于相关/不相关的二元评判。累计增益(CG)** ：计算位置1 到p的检索结果的相关度之和，，越大越好。rel描述文档相关性，据需求选择多个数值级别。但并未考虑文档位置，并不能体现靠前部分文档的质量。 **折损累计增益(DCG) ：**若搜索算法把相关度高的文档排在后面，则应该给予惩罚。 。后者采用指数，更突出相关性。由于其随着长度p单调非减，DCG与具体查询和结果列表长度有关，不利于不同算法间对比。**归一化折损累计增益(NDCG)**:将DCG除以理想结果(iDCG，即理想输出序列的DCG,用第一个公式):NDCG=DCG/iDCG。**多查询评价**：**MAP(Mean AP)：**对所有查询的未插值AP求算数平均。可以反映全部查询的综合效果，但在查询难度不平衡的条件下有误导。基于几何平均值的**GMAP**：基于几何平均值，提升相对强弱影响当各个查询间难度不均，或存在较难排序的主题时，GMAP更合适。**倒数排序（RR）**：第一个相关的文档的位置的倒数，越大越好。**平均倒数排序（MRR）**：对多个查询所得的倒数排序求评价(两个查询的第一个相关文档分别在位置2和4,则MRR=(1/2+1/4)/2=3/8,即平均在8/3位置找到第一个相关文档。预期倒数排序(ERR):当用户发现了第一篇相关文档后，后面的内容可能就不再关注了**。ERR为用户需求被满足时停止位置的倒数的期望**。**方差**：对于一个测试文档集合，检索系统常常对有的查询表现的很好，而对有的查询表现很差。通常情况下，一个检索系统对不同查询的方差，往往大于多个检索系统对相同查询的方差。 **结果多样化评价**：考虑多样性的原因：用户的单次搜索可能体现出多方面的需求，用户搜索可能存在歧义，需要展示多方面内容加以确认。基本形式：给定一个查询q，返回一个多样化的结果文档集合R(q)。R(q)应满足： R(q)中所有的结果文档都与查询q本身有较大的相关性。总体上要有较小的冗余度，以覆盖q的不同方面。 核心思想:降低用户无法获得所需信息的风险。两种衡量方式：隐式模型（只计算文档之间的差异性，不考量文档是什么内容）、显式模型（更加具体地考量文档所对应的用户意图，会从文档中抽取主题，并显式地实现主题的多样化）。隐式模型：最大边界相关性（MMR）：前一半表示文档集与查询 q 的相似性，后一半表示文档之间的多样性，lamda二者调节比例，文档用tf-idf或word2vec表征。显式模型：FM-LDA：考虑文档的不同子主题：表示文档Di包含主题Fj的概率，Yi=1表示某文档被选中。由此，**1-联乘部分**表示主题Fj至少被一个文档所涵盖的概率。当有约束时，返回的文档数量被限定为前篇。

**Ch9实体识别**

**信息抽取：**从语料中抽取指定的事件、事实等信息，形成结构化的数据。内容：8字方针："抽取实体，确定关系"。五类基本任务：命名实体NE(实体抽取)、模板元素TE(属性抽取)、共指关系CR(不同命名实体表达的相同含义)、模板关系TR(关系抽取)、场景模板ST(事件抽取)

**知识图谱：雏形理论**：1.**本体论**(本体：形式化的，对于共享概念体系的明确而又详细的说明，指特定领域之中存在的对象类型或概念，及其属性和相互关系；五元组表示法O={C,R,F,A,I})、2.**语义网络**(一种以有向图结构表达人类知识构造的形式)。3.**语义网**(通过给万维网上的文档 （如HTML、 XML文档）添加能够被计算机所理解的语义“元数据”，从而使整个互联网成为一个通用的信息交换媒介,语义网侧重信息的语义内容，并考虑计算机对文本内容的理解以及他们之间的相互交流)。**知识图谱优点：**1.找到最想要的信息：不再需要用户自行浏览、阅读和总结，而将信息直接呈现.2.提供最全面的摘要：对搜索对象进行总结，使得

用户获得更完整的信息和关联.3.让搜索更有深度和广度：构建完整知识体系，使用户获得意想不到的新发现。**知识图谱的基本形式**：由结点和结点之间的边组成，结点表示概念(或实体)，边表示关系(或属性，关系侧重关联，属性侧重特征)。数学上表现为一个有向图。点和边组成**知识图谱的基本单位**：三元组(实体-关系/属性-实体)。**知识图谱相关应用**：语义搜索(通过建立事物之间的联系，实现更准确、更直接、更完整的搜索);问答系统(在理解用户意图的基础上，将知识图谱作为大脑，基于推理能力提升交互体验);推荐系统(利用知识图谱提高推荐系统的推荐多样性和可解释性，提升推荐性能,如知识图谱上的游走，基于路径推荐)。**事理图谱：**描绘的是一个逻辑社会，研究对象是谓词性事件及其内外联系。借助图谱中的事理逻辑链接,可以形成对于事件的推理。**多模态知识图谱:**属性多模态、实体多模态。**命名实体识别**：识别出专有名称和有意义的日期等，并归类。是信息抽取中的核心任务。**两个子任务**：判别实体边界；判别实体类型。分类：3大7小：实体类(人、地、机构名)，时间类(日期、时间)，数值类(货币、百分比)。不是命名实体的例子：普通名词(飞机,公司)，非时间、日期、货币、百分比的数字，人的团体名称以及以人命名的法律、奖项等(如共和国、诺贝尔奖)。实体识别难点：不断有新的命名实体涌现；存在严重歧义(May名字or月份)；构成结构复杂(缩略词,别名,音译)；类型多样(共指关系，两个词说的同一个人)性能评价：Precision/Recall/F-value(分子为返回的正确答案数量 或+ ½的部分正确答案数量，如类型正确边界错误)**实体识别方法**：**基于词典(常作为基准算法)**：预先构建命名实体词典(来自领域公开数据)，出现的即识别为命名实体(简单快速,与语境无关,易部署和更新//难以枚举所有,词典构建和维护代价大，难以处理歧义)。**基于规则：**手工构造规则模板，对符合规则的实体进行识别，选用特征包括统计信息、标点、关键字等，以模式和字符串相匹配为主要手段。(优：规则能精确反映语言现象时性能好//缺:规则依赖于具体语言、领域和文本风格；代价太大、系统建设周期长、移植性差。需建立不同领域知识库)。**基于统计：1. 基于分类**（将标签当作类别，如8种实体，每种对应BMES四种类型，加上Other共33类标签；特征则选取当前词及前后各两词，抽取词性、字符类型、单词特征，总共15维。但要设计大量相关特征）。**词性标注问题：**词性是词汇基本的语法属性，通常也称为词类。词性标注就是在给定句子中判定每个词的语法范畴，确定其词性并加以标注的过程。词性标注问题，尤其是中文词性标注问题，面临一些困难和挑战：汉语是一种缺乏词形态变化的语言，无法从单词形态上来判别；常用词兼类现象严重，例如：科学技术（名词） / 这不科学（形容词）。**词性标注的方法：**与实体识别、分词类似:基于规则(人工或通过大规模语料学习规则，核心思想是按兼类词搭配关系和上下文语境建造词类消歧规则)；基于统计模型(HMM等面向词序列)；统计与规则相结合(先基于规则排除明显歧义，再基于统计模型标注，最后人工校验)。**2.基于序列模型：**与分词中的序列标注方法思路类似，区别在于标注的不同。针对命名实体的类别不同，引入了更多、更细致的标签种类。常用模型亦采用HMM、 CRF以及各种序列深度学习方法(如LSTM)等。**基于统计的方法的特点：**对特征选取的要求较高，需要从文本中选择对NER有影响的特征来构建特征向量（尤其是早期工作，深度学习技术发展后相对要求降低）；通常做法是对训练语料所包含的语言信息进行统计和分析，从中挖掘出特征；对语料的依赖也较大，目前缺少通用的大规模语料(对深度学习技术影响尤甚，特定专业领域影响最为明显);大部分技术仍需要进行人工标注训练数据。**基于统计方法的改进：**引入更多信息和领域知识以提升效果(英文词根词缀，汉字部首) **实体对齐/实体匹配**：对于异构数据源知识库中的实体，找出属于现实世界中的同一实体(利用实体属性信息,语义信息,实体间关系来判定，甚至是多模态特性(多模态知识图谱里的实体对齐))。 **实体消歧**：一个单词多个意思。抽取各个含义的相关内容(如描述文本)，建立关键词表；通过对关键词表的语义分析，从中抽取和归并相应的“概念”，然后对关键词进行语义表征，得到不同语义的表征向量。由此可以通过语义相似性（如余弦相似度）判断究竟属于哪种语义的实体。

**Ch10关系抽取**

**模板关系TR(关系抽取)**：实体之间的各种关系，又称为事实。通过关系抽取，将实体关联起来(职务、雇佣关系等)。**关系抽取任务**即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。意义：1.搜索引擎发现和关联知识的重要渠道(鸡腿和减肥)2.知识库构建与知识关联的基础性手段(在理解关系的基础上建立结构化知识库，为关联和推理奠定基础)3.是支持问答系统、推荐系统等应用的有力工具(准确、多样、有理有据)。**表达：**二元组(特定领域关系，类似二部图)，三元组(多类型关系，如企业间的商业关系)，多元组(时态关系)**关系抽取方法：**基于规则(手工定制规则，模式匹配)：通常针对特定领域的特定关系(据其特点设计针对性规则，较好准确率)；需要领域专家构筑大规模的知识库，代价大；移植到其他领域十分困难

基于模式：首先由种子关系生成关系模式，然后基于关系模式抽取新的关系，得到新关系后，从中选择可信度高的关系作为新种子，再寻找新的模式和新的关系，不断迭代直到无新关系或模型产生。法1：**DIPRE(双重迭代模式关系抽取)：基本元素**：元组(表示关系实例)，模式(包含常量和变量)。**基本假设**：三元组往往广泛存在于各个网页源中；三元组的各个部分往往在位置上是接近的；在表示这些三元组时，存在着某种重复的“模式。**DIPRE则考虑模式和三元组实例间的双重影响关系**(相比于通过检查部分网站来获取潜在的“模式”，并利用正则表达式来描述这种未必适用于其他网站、不可能穷尽所有模式的方法)，既要找到符合模式的三元组，也要找到用于生成三元组的模式。**算法流程：**1. 输入一组种子三元组实例R。2. 基于种子实例集合R，找到这些三元组在网页中出现的内容O (Occurrence)，注意寻找的时候保留上下文信息。3. 基于O，生成模式P。4. 基于P找更多三元组实例R，然后停止或返回第2步 。

**DIPRE生成模式步骤**：1. 将Occurrence归纳为Order(元素的顺序)和Middle(中间部分)。2. 定义模式: 模式的Order和Middle为 Occurrence 集合的Order和Middle，模式的URLPrefix、Prefix、Suffix分别为Occurrence集合中最长的公共URL前缀与前、后缀，其他部分采用通配符填充。 法2：**Snowball：仅信任支持度和置信度较高的模式，从而保证模式质量**。支持度(满足每个模式的三元组的数量),置信度（），即符合该模式的三元组确实符合相应关系的概率。**基于模式方法的优缺点**：不同算法的差异主要在于模式生成方法和匹配方法；适合某种特定具体关系抽取；基于字面匹配，没有引入深层次信息(词性,句法,语义)；移植性差，要为每个关系生成识别模式。基于机器学习：变换为分类问题：基于特征(难以找到合适的语义特征，如词性、句法分析树、依存树，而且对缺少NLP处理工具的和资源的语言，无法提取特征->考虑引入深度学习)/基于核函数(避免维度灾难，但参数难以确定)。**开放关系抽取：**预先定义的模板关系无法涵盖。考虑1.**基于知识监督**(基于Wiki等结构化知识库，从文本抽取关系信息)。2.**基于句法:**识别表达语义关系的短语, 可通过对于关系短语的句法结构约束来抽取关系;优点： 无需预先定义关系类别;缺点：语义没有归一化，同一关系存在多种不同表达(总部位于、设置于)。**远程监督：** 思想：如果某个实体对之间具有某种关系，那么所有包含这个实体对的句子都是用于描述这种关系，所以将其打包后训练模型再用于判断更多实体对间关系。 **局限：**语义漂移，不是所有包含该实体对的句子都表达该关系，迭代放大错误(考虑人工校验)。**优化方案：**动态转移矩阵(描述各个类间相互标错的概率)；规则学习(设计否定模式列表来去除错误标签)；注意力机制(不同句子对判别模型的贡献度不同)

**关系补全与链接预测：**已知图结构中的节点和部分边，推测其他可能存在的边[现有图谱规模稀疏，人工方法成本高，新关系不断产生]**事件抽取**：事件是(5W1H)相互作用所产生的客观事实；句子级别；推理任务的前提。信息抽取的基本任务：场景模板ST(事件抽取)。**事件基本要素：**事件触发词(核心词),事件类型,事件元素,事件元素角色。事件触发词的检测与分类是事件抽取的基本任务。**事件抽取的模板**：通过模板元素TE，通过事件元素识别，将元素填入模板相应**槽**内即可。**限定域事件抽取**：预先定义好目标事件的类型及每种类型的具体结构**.开放域事件抽取**:无监督(聚类);知识监督. 以事件代替实体作为节点，可以将知识图谱拓展为**事理图谱**

**Ch11数据准备 数据挖掘基本含义：**从海量数据中提取或挖掘潜在的知识和规律，用于支持当前的判断或未来的决策：1.数据准备：筛选、清理并整合有待挖掘的数据；2.数据建模：使用智能方法建模数据并提取规律；3.知识表示：以用户能理解的方式展示所得知识。**数据挖掘目的：**揭示满足以下条件的模式（Pattern）或模型（Model）1.有效性：在处理新数据时具有足够的可信度；2.实用性：可以用于实际的项目并产生指导价值 3.解释性：能够被人们以经验或领域知识所理解.4.意外性：所得结果并不是那么直观和显而易见(把握尺度的重要，既要有一定深度，又不能过于颠覆)。**网络数据挖掘：**万维网里的数据挖掘(对领域知识要求没那么高，但结构复杂、规模宏大、时效性强) **数据挖掘方法：**1. 分类：有监督。2. 聚类：无监督。3.关联规则。4. 离群检测(异常检测)。**关联规则：**事务性数据：一条记录对应一个项目的集合（无序）**关联规则：**旨在分析事务型数据，从而根据一部分项目的存在记录来判断另一部分项目是否同时存在于事务中(表示相关性而非因果性)。**基本形式**：，AB均为集合。基本指标：1. 支持度：{A+B}在全体事务中的比重；2. 置信度：{A+B}占A出现的事务中的比例。频繁项集：支持度高于阈值的项目集合A。基本方法是穷举所有可能集合计算其支持度，复杂度过高。**改进方法1：Apriori**：先验原理：如果一个项集是频繁的，那么它的所有子集也是频繁的。（非频繁的项集，其所有超集也是非频繁）算法：逐步减去所有非频繁项集，基于频繁项集生成其超集。（1频繁项集每个集合含1个项，删去支持度小于阈值的，留下的项生成2频繁项集，以此类推）,减少了对候选项集的遍历，但仍是“生成-测试”范式。**法2：FP-Growth**:进一步减少空间开支，直接从结构中提取频繁项集，本质是输入数据的压缩表示，通过逐个读入事务，并将事务映射到FP树的某条路径来构造。**建树过程：1.**先扫描一遍数据集，得到频繁项为1的项目集及其出现次数，定义最小支持度（项目出现最少次数），删除那些小于最小支持度的项目，然后对1项目集按出现次数降序排序。**2.** 根据项目（物品）出现的次数重新调整物品清单(如物品清单第一条为I1,I2，但是实际I2的1项目集出现次数更多，所以修改为I2,I1,其余每一行类似),将排序后的项集(每一条物品清单)逐步读入并建立树状结构(对出现相同的节点进行累加，如I2:2)；对相同项节点采用指针连接，方便快速访问。**基于FP树生成频繁项集（自底向上）**：**1.**查找以e结尾的频繁项集：判断e是否为频繁项集，将e的前缀路径转化为条件FP树（更新路径上支持度计数，删除非频繁的项）,对于该树中每一项x生成二项集{xe}。**2.**考虑更长结尾如以{de}结尾的频繁项集的子问题：在前一张图上统计与d相关的支持度求和(支持度为2，为频繁项集)，以de为结尾得前缀路径，生成条件FP树，删去非频繁项，得到ade支持度为2，也频繁。此时FP树为单路径，所以停止递归，即不再看ade为结尾的条件FP树；再用ce结尾生成条件FP树，为null，所以无更多频繁项集。 3.迭代完后，再回到第一步，选取倒数第2的点d重复e的操作，依次类推

**异常检测：**旨在发现与大部分其他对象不同的数据。异常数据错误数据，是包含不同寻常规律的数据。异常是相对的。常见应用场景：欺诈检测（银行异常交易）、智慧医疗、入侵检测（网络异常行为）方法：1. （基础）基于分布：统计方法，前提是识别数据集的具体分布。基础方法：基于一元正态分布的离群点判定。已知参数的前提下根据正态分布判定离群概率，将一元正态分布扩展到多元。2. （进阶）基于度量：基于距离或密度，如计算数据点到K近邻平均距离、采用K近邻距离的倒数作为密度、给定半径内点的个数。3.（更进阶）基于聚类：聚类分析可发现强相关数据集合，不相关即为离群点。如抛弃远离其他簇的小簇，但簇个数影响结果；聚类所有对象，评估对象属于簇的程度，指标：点到簇中心的距离、相对距离（即与所有点到中心距离中位数之比）**数据预处理** 常见数据质量问题：数据测量/采集等过程出现的错误、噪声、离群点(对单链聚类有很大影响，使不该聚在一起的聚在一起)、缺失值(由于数据采集不完整或者部分属性值在部分数据中不适用(儿童无收入)；采取**删除和填补**并重的方法)、重复数据(多源数据归并导致，基于数据整合方法(实体对齐任务))。**常见预处理方法：**数据整合、数据采样、维度规约、数据离散。**数据聚合：**将两个或多个对象合并为单个对象。1. 目的：归并多个数据源的数据到同一格式下，在一定程度上解决重复数据问题。2. 动机：减少时空开支，对象群的属性比个体更稳定。3. 多源数据归并可能的问题：不同属性名称、不同单位或尺度、不同属性统计方式、统计上的不一致。**数据采样：**选择一部分数据对象的子集进行分析的常用方法。1. 目的：采取小规模样本起到近似效果，降低开支；在要求精确时，采集小规模样本初步分析了解数据特性。2. 问题：采样缺乏代表性会产生误导（采样数据应在统计指标上近似原数据，如均值方差）.3. 常用方法：**简单随机采样**(等概率采样所有对象：**无放回、有放回采样**，两种方法无本质区别，后者更简单(概率不变))、**分层采样**。4. 影响采样效果因素：**样本容量/采样规模**，较大的样本容量更能完整代表数据，但降低了采样的收益；较小的样本容量在采样收益上更高，但可能造成信息的损失。**启发式采样规模确定**：分组采样（分为10组数据，组内相似，不同组差异大。样本容量需满足能从每个组至少取到一个点的概率>=90%）。**不良倾向**：采样不应该有任何倾向性。**维度规约：**数据集中可能涉及大量特征，通过维度规约删除不具有区分度的特征，同时可能降低噪声，在避免维度灾难的同时，模型更易理解易于可视化。1**.维度灾难**：维度增加，数据分析困难程度大幅上升；可能由于：1.计算量指数级增长而难以处理,2.数据稀疏、无足够数据可建模(可能导致过拟合训练样本仅能反映部分属性的规律)。2.**代表方法：主成分分析**3.另一思路：**特征子集选择**：使用特征的子集而非归纳新特征，目的在于去除冗余特征和不相关特征。除删去多余特征，也可为特征赋予不同权值。4.**信息损失**：维度规约可能造成信息损失，甚至产生误导。

**数据离散化(主要以分类问题为例)**：将连续属性转化为分类属性，eg:Pointwise的有序回归方法，根据回归值所在区间决定文档标签。1. 二元化：离散化的基本形式，将连续或离散属性转化为一个或多个二元属性(根据属性的各个取值，进行二进制编码000,001,010等，或者one-hot，关联问题倾向后者，因为每个item独立)。2. 非监督离散化：用于分类的离散化方法根本区别在于其离散化过程是否有监督（是否使用类别信息）。非监督常见方法(根据数据本身特性)：等宽(区间间隔相同)、等频率(每个区间样本数量均衡)、等深、K均值。3. 有监督：注重问题导向，目的在于取得更好的效果。**基于熵**，区间i的熵，j为某个属性值，p为占比，总熵为加权和。熵越小区间内纯度越高（标签越一致）。一种做法是先二分，选择熵最小的点分割，对其中较大熵得部分再进行分割，以此类推。**主成分分析PCA**思路:通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的变量为主成分。通过这种方式，可以采用较少的综合指标综合先前存在于各个属性（且相关）中的各类信息，而综合指标之间彼此不相关。二维特征例子：把数据点圈进一个椭圆，椭圆短轴方向数据变化较少区分度低，删去短轴，将坐标轴变换与长轴平行(点投影到长轴上)。高维类似，找出几个最长的轴作为新维度。选定轴（坐标系）：**最大特征值对应的特征向量可以最大化投影方差**。所以求数据样本对应**协方差矩阵**的最大的K个特征值，其特征向量对应的线性组合形成K个新的综合指标。K个特征值的比重反映了主成分的信息量，一般应大于0.85。注意：1.PCA依赖于原始变量，也只能反映原始变量的信息，原始变量的选择很重要。2. PCA内在假设之一：原始变量直接存在一定关联性，如果原始变量本质上相互独立降维可能失败(数据越相关，降维效果越好)。3.PCA结果未必清晰可解释，与选取的原始变量及数据质量等相关(因为新生成的属性是原属性的线性组合)。

**Ch12分类**

基于规则的分类：就是一组if-then来分类，规则->标签。基于规则的分类器所产生的规则集，应具有两个重要性质：互斥原理：如果规则集中不存在两条规则被同一条记录触发，则规则是互斥的。该性质确保每条记录**至多**被一条规则所覆盖 ，发生矛盾的解决方案：重新制定规则、**采取有序规则(**只满足第一个触发的规则，对规则集按优先级降序排列。常见的规则排序方法：1.基于规则的排序方案：按照规则质量如准确性进行排序；2.基于类的排序方案：同类规则排在一起，相对顺序被忽略(但不同类间的排序还需考虑))、采用投票机制。穷举原理：对于属性值的任一组合，规则集中都存在某条规则加以覆盖。该性质确保每条记录**至少**被一条规则所覆盖，有无法被分类的项时：制定兜底规则。**如何制定规则分类器：1.最基础的方法：人工制定规则进行分类**(相较于纯手工打造，效率更高，运用更广泛,如持续不断维护更新的敏感词列表,如果人工规则得到定期更新，效果往往较好，然而，维护规则成本较高，需要持续更新)。**2.利用算法自动生成规则：直接方法：**从数据中自动学习规则，如顺序覆盖、RIPPER、CN2等算法。**间接算法**：借助其他分类模型学习规则，如从决策树中提炼规则。**顺序覆盖**：采用某种评估以贪心方式进行学习，算法基本流程（二分类）：1.算法开始时，决策表为空，即不包含任何规则；2.每一步针对某个类 y，提取覆盖当前训练集的最佳规则(什么是好的规则？覆盖的样本尽可能的多，同时样本类别尽可能一致)；3.如果规则能覆盖大多数正例（即 y），而没有或覆盖极少负例（非 y），则保留该规则(将该规则加入决策表的尾端，同时删除该规则覆盖的所有训练样本)；4.重复上述过程，直至满足终止条件（例如：某个增益的阈值，如熵或准确率）。**规则增长策略**：搜索空间呈指数增长，找到一个最佳规则的计算开支很大，所以一般而言，采用一种贪心的增长规则来解决指数搜索问题，常见的分类规则增长策略包括“一般到特殊”和“特殊到一般”两种：**前者**初始规则条件为空，给定目标标签 y，逐步加入合取项（通过AND运算相连）来提高规则质量；**后者**随机选择一个正样本作为初始种子(保留其所有规则)，逐步删除规则中的合取项，来覆盖更多的同类别正例。**顺序覆盖里为何要删除样本：**顺序覆盖的本质是一种贪心法思想，新规则对于整体分类的增益具有边际效益，若不删去则会影响后续新规则准确率的评估。**基于决策树生成规则的间接方法：**决策树从根节点到叶节点的每一条路径都对应一条分类规则，路径中条件为规则前提，叶结点标号为分类结果，满足互斥和穷举。**基于规则分类器的优点：**基于规则的分类器具有较好的可解释性和直观性；基于规则的分类器生成较为简便，分类也较为迅速(如果面临违反互斥原则的情况，可能涉及多条规则投票，相对较为复杂)。**基于监督学习的分类：决策树**：一种典型的分类方法，首先对数据进行处理，利用归纳算法生成可读的规则和决策树，然后使用决策对新数据进行分析。本质上，决策树仍是通过一系列规则对数据进行分类的过程。决策树的生成： 1.特征选择：选取对于训练数据有着较强区分能力的特征。2.生成决策树：基于选定的特征，逐步生成完整的决策树。3.决策树剪枝：简化部分枝干，避免过拟合因素影响。决策树的特征选择问题：选取对训练数据具有区分能力的特征，从而提高决策树学习的效率。如果某特征分类的结果与随机结果没有很大的差别，则称这个特征是没有分类能力的。经验上，扔掉这样的特征对决策树学习的精度影响不大。信息熵：熵是表示随机变量不确定性的度量，不确定性越高，熵越高。设D是一个取自有限个值的离散随机变量，其概率分布如下：，则随机变量的熵定义为：。对于二分类时，D取值仅0/1，则p\_0=0或1时，没有不确定性；p\_0=0.5时，熵取值最大，随机变量不确定性最大。特征选择准则:1.信息增益:特征A对训练数据集D的信息增益，定义为集合D的经验熵与特征A在给定条件下对D的经验条件熵之差，即:，其中，公式的后半部分，也就是经验条件熵的部分。**信息增益越大，说明特征越好**。2.信息增益率：信息增益虽然能够较好地体现某个特征在降低信息不确定性方面的贡献，信息增益越大，说明信息纯度提升越快，最后结果的不确定性越低。但是，信息增益也具有一定的局限性，尤其体现在更偏好可取值较多的属性，取值较多，不确定性相对更低，因此得到的熵偏低。信息增益率：,**增益率越大，特征越好**。本质上是引入一个惩罚项，取值越少，惩罚项越小。其中：。信息增益率有矫枉过正的危险，采用信息增益率的情况下，往往倾向于选择取值较少的特征。当取值较少时，较小，因此惩罚项相对较小。**目前通常采用折中的方法**：先从候选特征中，找到信息增益高于平均水平的集合。再从这一集合中，找到信息增益率最大的特征。3.基尼指数：**基尼指数越低越好**，表明随机样本被分错的概率越低，相应的信息纯度也就越高。假设有K个类，样本点属于第k类的概率为，则基尼值为，越小表示集合纯度越高。属性a的基尼指数，V为a划分的集合数目，即a取值种类数。**决策树的最终目标**：在于使每个节点所对应的样本类别均为“纯”的。决策树的剪枝：剪枝的原因在于训练过程的“过拟合”问题（如果训练集与测试集效果都不好，说明出现“欠拟合”现象。如果训练集效果好，而测试集效果不好，说明出现“过拟合”现象），过拟合出现的原因在于训练过程中过度迁就训练数据特性，而导致构造出过于复杂、过于细枝末节的决策树，泛化能力较差(可能由于训练集本身的特有样本导致了干扰，无法代表测试数据的分布)。解决这一问题的办法在于对已生成的决策树进行简化，即“剪枝”。1.预剪枝：在生成决策树的过程中即进行剪枝。每个节点划分前，衡量当前节点的划分能否提高决策树的泛化能力。2.后剪枝：在生成决策树之后再进行剪枝，称作“后剪枝”。自底向上考察每个非叶子节点，考虑将该节点替换成叶子节点后能否提高泛化性能。3.基于损失函数剪枝：通过极小化决策树整体的损失函数来实现剪枝。剪枝的过程是在已生成决策树之后进行的，也属于“后剪枝”的一种。树T的损失函数基本形式形如：，其中，为模型复杂度（节点个数），可定义为。该部分用于表示预测误差，表示第t个节点上第k类的样本数。调节两者影响，较小的对应较复杂但拟合度较好的树。**K-最近邻分类（KNN）：**表征空间上相近的文档应属于同一个类别。**模型输入：**训练样本集合与距离度量 ;K值，用于限定最近邻的数量。**算法流程：**给定未知样本，首先计算与其他样本的距离，找到K-最近邻。基于K-最近邻的类别确定分类结果（投票法）。**影响KNN效果的关键问题：1.距离度量：**对于高维空间而言，最基本的度量方式为欧式距离：。如果采用0/1向量，则可使用汉明距离代替，相当于统计有多少维的数字不相同，该数值即为汉明距离。除此之外，对文本而言（如采用TF-IDF），可使用余弦相似度。其他可采用的度量如马氏距离、无穷范数（向量最大值）等。**2.K的取值：**K太小，则容易受噪声干扰，K太大，又可能导致错误涵盖其他类别样本。一般K取奇数，保证投票法有结果。采用倒排索引来寻找K-近邻：对于网页文档来说，寻找K-近邻的过程，可以近似视作基于给定查询，去寻找K个最佳查询结果（文档）的过程。将元素视为⽂档，元素的每⼀维视为词，坐标视为不同词项的频率即可，这样同ch4一样构建词典和倒排表。（利用查询中的词项对文档进行初筛，只有具有某些相邻词的文档才保留下来作为备选）最近邻分类的特点：1、最近邻分类是一种典型的基于实例的学习，使用具体实例进行预测，而不需要对数据进行抽象（如提取特征）。2、最近邻分类是一种消极学习，不需要模型，但分类过程开销很大。相比之下，积极学习方法训练模型较为费时费力，但基于模型分类很快。3、最近邻分类基于局部信息进行判别，受噪声影响很大。4、最近邻分类需要慎重选择度量并预处理数据，否则可能被误导。例如，借助身高体重进行分类，身高波动范围不大，而体重差距巨大。**支持向量机SVM**：如果仅考虑二分类问题，那么分类问题可以转化为寻找一个超平面，实现对于高维空间中的节点进行有效分割。应该选择“正中”的最大间隔超平面，容忍性好，泛化能力强，符合这样条件的超平面，在线性可分的条件下“存在且唯一”。超平面方程：w^T \* x +b=0。w,b 为参数， w向量方向垂直于超平面，离超平面最近的节点被称作“支持向量”。间隔r=2/ ||w||。**求解：**该形式符合凸二次规划问题特征，可以借助拉格朗日对偶性，通过求解对偶问题加以求解；其拉格朗日函数的下确界为原凸规划问题最优解的一个下界，所以求解该问题可以获得原凸规划问题的最优解。先L(w,b,α)对w和b求偏导，得到的结果再回代到L，得到一个新的形式->再由SMO序列最小优化算法求解(基本思路：不断执行如下两个步骤直至收敛:第一步：选定以对需要更新的向量αi 和 αj；第二步：固定αi 和 αj以外的参数，求解对偶问题以更新αi 和 αj。挑选αi 和 αj时，注意挑选差别较大的一对以尽快提升目标函数)。**线性不可分问题：**将样本映射到一个更高维的特征空间，使得在这个特征空间线性可分。**核函数：**某些样本在低维空间时线性不可分，通过非线性映射将其映射到高维空间的时候则线性可分，但非线性映射的形式、参数等难以确定。核函数的目的，在于将高维空间下的内积运算转化为低维空间下的核函数计算，从而避免高维空间可能遇到的“维度灾难”问题。如何选择核函数：穷举法：一个个试过来，选择效果最好的一种；混合法：将多个不同的核函数混合起来使用。（线性核函数与高斯核函数（径向基）是最为常用的）。**软间隔问题：允许少数样本不满足超平面约束，因为**有时可能既没有办法找到一个线性可分的超平面，也无法确定合适的核函数使得训练样本线性可分，也有过拟合的可能。引入惩罚项即可。**SVM与多分类问题：**SVM是标准二分类器。常见策略：一对多（一个SVM针对一类，即把一类与其他所有类区分开），一对一（一个SVM区分两类，划分出的两类再分别由另外两个svm划分）**不平衡分类问题：**具有不平衡分布的数据集在日常应用中极为常见，如果依赖于效果指标优化分类结果，在不平衡分类条件下可能被误导。不同应用场景中，对于不同类别样本的分类需求与指标侧重也不同(邮件分类：（对垃圾邮件的）召回率，保证较高准确率；智慧医疗：牺牲（对病人判断的）准确率，保证较高召回率)。不平衡分类问题解决方案：（1）代价敏感学习：引入代价矩阵，衡量将一个类错分到另一个类的代价。例如，宁可多判断一些疑似患者，不能漏掉一个病人的情况下，可将 “病人”错分至“健康”的代价设为100，而正常人错判为”有病”的代价为1 。最终优化目标由原先的准确/召回变更为加权后的代价。（2）抽样方法：基于采样，通过改变样本分布来缓解不平衡问题。常见的策略有以下三种：过采样：提升少数类的样本比例，例如从少数类中进行重复随机采样(有放回，对KNN好用，利于投票)。欠采样：降低多数类的样本比例，例如用多数类的采样样本而非全样本(采样要具有代表性)。混合采样：同时采用前两种策略。也可以利用已有的少数类样本，通过K-最近邻生成新样本，采样中要注意噪声问题，噪声也可能被复制多次。

**Ch13聚类**

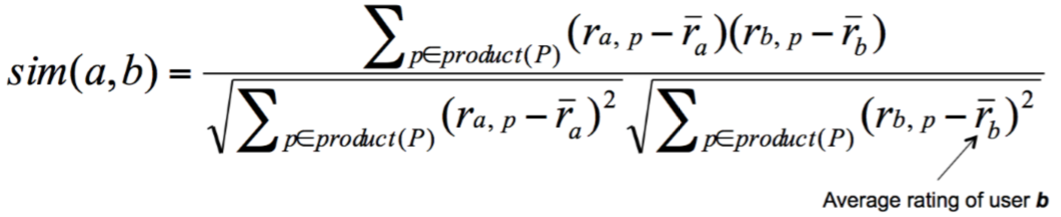
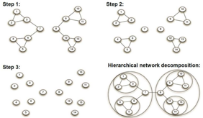
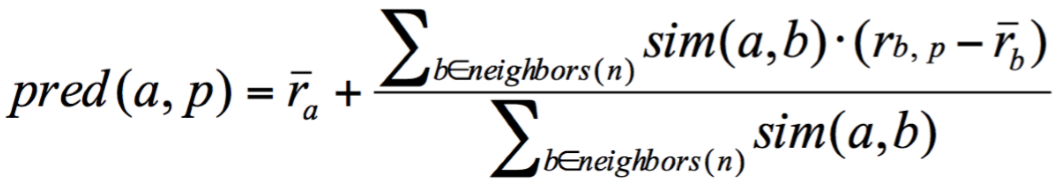
**基本概念：**簇：将样本表征为高维向量，相似样本将自发地形成“簇”的结构。簇内相似（距离较近），簇间相异（距离较远）。**聚类的目的**：将样本分为若干个簇，是无监督学习方法(借助相似性度量自动生成)；聚类依据具有主观性，其合理性受问题与方法定义的影响，存在一定不确定性。**聚类的三个基本问题：1.在数据中，样本是如何形成簇/社团的？**即样本的“群体性”的依据是什么->有主观性，取决于聚类依据/目的的选择。**2.如何度量样本之间的相似性？**不同的相似性度量可能导致截然不同的簇。相似性度量通常根据算法或表征加以选择，例如大多数时候采用距离衡量相似性，但有隐患(向量表征人的爱好，但好友间向量未必就相似) **3.簇的数量如何确定？**由于聚类没有天然标签，簇的数量往往是个开放性问题，过少(太笼统)或过多(过拟合、噪声点)的簇都应该避免，这会导致失去代表性，但未必能通过簇数调节。三类问题：内容(相似性文档归类)、用户(相似条目的关联)、结构(网页簇挖掘、社团挖掘)。需要区分：分类问题(有监督学习)、简单切分、排序(仅分组，没有按表征和相似性归类)、基于Query查询(按照外部条件匹配而非自身数据)。**不同簇类型**：**明显分离的簇**(每个簇中的点到同簇中其他点的距离要小于到不同簇中所有点的距离，可有效实现聚类)；**基于中心的簇**(每个簇中的点到簇中心的距离要小于到其他簇中心的距离，簇中心的选择将严重影响簇的形成。未必完全分离，但根据中心可有效分割，一般趋于球形)；**基于邻近的簇**（每个簇中的点，与其最近邻在同一个簇中。依赖最近邻，对于不规则或缠绕情况下的聚类有奇效，但易受噪声影响）。**基于密度的簇**(簇是对象的稠密区域,簇的含义相当于被低密度区域分开的高密度区域，不仅适用于不规则形状，甚至可解决噪声或离群点问题)；**基于概念的簇**(簇的含义是某种具有共同性质的对象的集合，对象的性质由概念定义，一般较为复杂，常见于模式识别领域)。**常见聚类方法：**分类：层次聚类：树状，除叶节点，每个节点是子节点的并集；叶节点一般为单样本组成的单元簇 。划分聚类：每个对象属于一个簇；簇之间不重叠。其他区分规则：**重叠(非互斥)聚类**中，每个样本可能属于多个簇(对应 多标签分类)。**模糊聚类**中，每个样本以一定从属度隶属不同聚类，从属度之和应为1(归一化要求)。簇的同构与异构：不同的簇可能具有不同的规模、形状、密度等属性**。K均值聚类：**簇的中心近似反映整个簇的属性。通过设定K个中心，来形成K个簇，然后不断更新簇中心的向量，来更新聚类的结果，直至收敛。**选定初始中心→聚类(将每个样本分配至最近的中心所在簇)→更新中心位置(基于划分出来的每个簇内样本向量均值)→聚类→更新中心位置**→...(至收敛，即聚类结果不再更新) 。**若干细节：1.**簇中心由簇中所有样本向量的算术平均决定；2. K均值的初始中心往往随机选定，而不同中心可能导致不同的聚类结果。3.大多数情况下，聚类过程将在少数几轮迭代后收敛（即不再更新）。也可将停止条件修正为“低于一定数量的节点变更簇的归属”。4.K均值聚类的**复杂度为O( n \* K \* I \* d )**，与样本数量（n），簇数量（K），迭代次数（I），向量维度（d）相关。5.样本归属可由欧氏距离或余弦相似度等计算。**聚类效果衡量：**平方误差和(SSE)：，x为样本，m\_i为簇C\_i中心，x属于簇C\_i。SSE越小越好，评价不同聚类方法时，在同等K下比较SSE(K增加,SSE趋于下降)。簇中心为簇内样本均值时，该簇的SSE\_i为最小，所以和K-means的更新中心位置是一致的。初始中心的选择：1、采少数样本，借助其他聚类先确定出初始中心(如层次聚类，但开销大，所以只取少数样本先试试，而后用K-means)，仅适用于K较小的情况。2.选择多于K个中心，然后从中挑选分隔较为明显的。3."**后处理"**：聚类后修补生成的簇。清除离群点(小的)簇；拆分松弛(SSE高)的簇；合并紧凑(SSE低)的簇；引入新的中心或打散一个簇后重新划分(新中心选择为距离所有簇中心最远的点)。4.二分K均值：先分为2个，然后不断选择其中一个分裂(选择的标准可以是样本数较大的簇，或者SSE较高的簇),此法受初始中心的影响不大。空簇处理：1.如果以样本作为初始中心，则不会出现这种情况（簇内至少一个点）。2.可以新生成一个簇来代替空簇(思路类似后处理)：选择一个最远样本点新生成一个簇；将最大SSE的簇拆分。离群点的处理：在使用SSE衡量聚类结果时，离群点可能造成负面影响。离群点可能影响簇的代表性，并且SSE也较高(这种情况下，应该删除离群点以保障聚类质量);但需注意离群点不是任何情况下都能删除（有时明显的离群点可能反而是我们研究的目标，如交易中的异常或欺诈行为；离群点的识别和处理可回顾离群分析部分）中心的增量更新：某个样本进行重新聚类时更新中心。要么不更新，更新必然是一对：好处在于不会有空簇的出现(单点簇必然维持原样)。缺点在于可能导致次序依赖性(不同次序导致结果不同)，同时开支更大K均值聚类局限性：当簇存在不同规模、密度及不规则形状的情况下结果较差；易受离群点干扰（解决：类似选初始中心思路，采用偏大K，进行合并）**层次聚类：**不需要预设簇的数量；随时根据需要调节：在相应的层数切分树状结构；结果往往可以对应到具有一定意义的分类学目录上。**凝聚式聚类（自下而上）：**（更常见）将所有样本视作个体簇，逐步合并最接近的两个。**邻近度矩阵**：存储两两簇之间的邻近度。**迭代流程：**合并邻近度最高的两个簇→基于更新的簇重新计算邻近度，更新邻近度矩阵。**不同凝聚式聚类方法的主要区别**在于不同的邻近度定义：单链 MIN：不同簇最近的点之间的邻近度，越近邻近度越高。擅长处理非椭圆形状的簇，但对噪声比较敏感。全链 MAX：不同簇最远的点之间的邻近度，越近邻近度越高。对噪声不太敏感，但可能使得较大的簇变得支离破碎。组平均：所有来自不同簇的两点之间的平均邻近度(计算代价大)。中心距离：来自不同簇的两个簇中心之间的邻近度(也可用合并两个簇导致的SSE增加值等度量方式)。**分裂式聚类（自上而下）：**从包含所有样本的完整簇开始，每一步分裂一个簇。二分K均值、最小生成树(MST)聚类：基于差异度矩阵生成最小生成树(节点间权值最小)→每步断开差异度最大的一条边，形成一个新簇。MST聚类和单链凝聚聚类结果相同。层次聚类的局限性：每一步合并都是最终的(一旦做出合并两个簇的决策，则无法撤销)；没有全局的优化目标函数(每一步都是局部最优而非全局最优); 不同的聚类方法（邻近度定义），或多或少都具有一些问题.**基于密度聚类：**只有达到一定密度，才足以成为一个簇(密度指一定半径的样本数量)。DBSACN算法：(**核心参数**：预设半径Eps与半径内样本数阈值MinPts)**三类样本点**：核心点(稠密部分内部的点，指定半径内样本数超过阈值的点)；边界点(非核心点，但是处于稠密区域边界内/上的点)；噪声点(处于稀疏区域的点)。基本流程：1.将所有节点区分为核心点、边界点或噪声点 2.删除噪声点 3.将所有距离在预定半径内的核心点之间连一条边 4. 连通的核心点形成一个簇 5.所有边界点指派到与之关联的核心点在的簇中。优点：对噪声较为鲁棒，且可以处理不规则形状。缺点：如果簇的密度变化很大，DBSCAN的效果可能会受到影响。参数选择：选择出现明显拐点的位置对应的距离，作为Eps的大概取值。**聚类问题的评估**

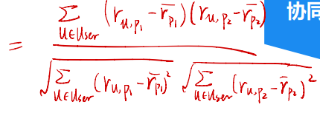
为什么需要评估聚类结果的“好”与“坏”：1.确定数据集的聚类趋势，确定是否真的有群体性2.确定合理的簇的个数3.比较两个簇，或者比较基于两种方法的聚类结果，看哪种结果更合适4.将聚类的簇与已知的客观信息（例如，外部提供的标签、 Query等）进行比较。**聚类问题评估的类别：非监督评估(内部评估)**：仅使用数据本身的特性**1、基于邻近度矩阵(基于密度的聚类方法不适用)：**最理想的聚类结果应该是：簇内的点邻近度全为1，簇之间的邻近度全为0，所以实际的邻近度矩阵与理想化矩阵的相似度，可以用于评估聚类结果。观察相似度矩阵是否出现对角模式，出现了则结果好**2、凝聚度与分离度：**簇内的邻近度定义为"凝聚度"，簇间的邻近度定义为"分离度"，凝聚度越高，分离度越低，聚类效果就越好。**基于图**，凝聚度由簇内各点邻近度之和定义，分离度由簇间各点的邻近度之和定义。**基于原型**，凝聚度由簇内各点到中心的邻近度之和定义，分离度由簇中心到其他簇内的中心（或点）的邻近度之和定义

**有监督评估(外部评估)**：引入外部信息，衡量聚类结构与外部结果的匹配程度。希望每个簇能完美对应到某个类别标签。**面向分类的度量：**熵(衡量每个簇由单个类的样本所组成的程度)、纯度（簇在多大程度上包含单个类的对象，以最多类的比例计算）、准确、召回、F值。**面向相似性的度量：**比较分类标签和聚类结果各自对应"理想"矩阵的相关性：

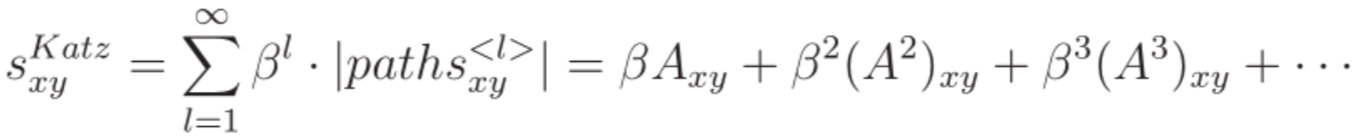
**相对评估：**比较两个簇或者两个聚类结果，往往借助前两种的度量手段。

**Ch14推荐系统**

******长尾现象、马太效应**(跟风，好的更好，弱的更弱)。**推荐系统的数学形式：**基本要素：用户集合X，物品集合S，评分集合R，效用函数u:X \* S->R ;矩阵角度看：M\*N维矩阵（矩阵补全问题）。**推荐系统三个关键问题**：**1. 收集“已知”的评分信息，生成效用矩阵（R）**：如何收集数据：显式数据(可直接获得的用户评价数据,往往难以获得。通过用户标注的方式获得，例如众包), 隐式数据(基于用户行为等信息，间接地判断用户的倾向性，如何保障负样本质量？)没有显式评分数据的情况如何处理？**2. 基于已知评分，推断未知评分**：着重关注高分元素（效用矩阵的严重稀疏性以及新用户/新物品的冷启动问题）**3. 推荐结果的评估：由对推荐问题的定义决定：（1）**用户评分预测，**均方根误差RMSE** ,r分别对应预测和真实评分。（2）视作分类问题(推荐正确与否)，完全二分类：利用Precision/Recall/F值评估；或排序后返回TopN，通过Pre@N,Rec@N评估(理论上限可能<1)。(3)排序问题，基于推荐物品受欢迎程度(预测评分)，用排序评估方式评估。**除此之外，更关注高分预测内容；且推荐应有多样性(避免茧房效应)**。**基于内容的推荐**：基本想法：用户的偏好一般相对稳定->给用户推荐其以前喜欢物品相类似的物品（偏保守）。**构建物品画像**(向量，电影元数据、文章关键词/主题模型/tf-idf等)**和用户画像**（基于其评分过的物品画像加权估计），用物品和用户的相似度来评分，可用余弦相似度。优点：1.每个人的推荐过程相互独立，不需要其他用户的数据（冷启动问题不那么严重）2.可以为具有独特偏好的用户进行有效推荐，不受大众倾向性和热度的影响。3.可以推荐新物品或非热门物品。4.推荐结果有着较好的可解释性，推荐的同时可列举内容特征作为推荐的依据。缺点：1.难以找到合适的特征，特别对于对于非结构化信息(如图像)；部分特征的提取可能存在误导性。2.难以给新用户推荐，如何建立新用户的用户画像？3.过度特化现象：永远只能给用户推荐局限于其画像中的内容；用户的多方面兴趣难以体现；难以通过他人的评价对推荐结果进行评估。**如何解决多样化问题**：引入多样性评估(最大边界相关性MMR)。**基于路径推荐：**知识图谱+游走(概率最大)。**基于协同过滤的推荐：**基本思想：其他用户的浏览行为对当前用户有借鉴作用，有相似爱好的用户在评分上也具有一定的相似性。推荐系统的本质就是矩阵补全问题。相应的，协同过滤的思想在于基于矩阵的其他行，协助填补本行的空缺。**基于内存的协同过滤：基于用户推荐的评分预测：**目的：找到相似用户，并基于这些用户的历史行为进行推荐。**最近邻的寻找过程**：历史评分行为越相似，用户之间未来行为的相似性就越高。基于共同评分的物品，衡量用户之间的相似性如下：注：计算相似度时，如果某个user-item对未评分，则对应r设为0，不用减去平均分。分子上是看共同评分的物品，分母是看各自有评分的物品。根据历史上其他用户对于该物品的评分，结合用户之间的相似性作为加权，预测评分结果。相似性计算与评分预测中，都通过减去平均值来抹去个人评分偏好的影响：有时只考虑K-最近邻的情况。公式中的neighbor(n)即缩小为K-最近邻的集合。**注意**：1、只考虑对指定Item有评分的邻居，无评分的跳过，不算在K近邻里。2、通常情况下，低于0的相关性可以被忽略-》K-最近邻集合可能实际上小于K个邻居。基于物品（Item-based）的推荐采用类似的设定。**基于物品推荐的评分预测**：衡量相似物品的标准，不是物品本身的属性，而是不同物品的评分历史。同一个人给两个物品打出相似分数，说明他认为两个物品相似；越多这样的人， 两个物品越相似。物品相似度计算与User-based类似，即构造两个向量，每维减去平均值，然后计算Pearson相关系数。如果未评分，则直接设为0，不用减去平均数。，**预测评分：**,r\_ix是x用户对i产品的评分，N(i:x)是x打过分的物品

实践中，基于物品的推荐效果更好:物品的属性相对单一，而用户的偏好则更为丰富多样。相应的，人们喜欢某样物品的理由也相对较为集中，而用户可能在不同情况下体现出不同的偏好。**基于内存的推荐的优缺点**：优点：可适用于任意种类物品，效果较好。不受多模态、非结构化信息表征与特征选取的困扰。缺点：冷启动：用户与物品都存在冷启动问题。稀疏性：用户评分记录严重稀疏，很难找到评价过同一物品的用户。热度偏差：更倾向于推荐热门物品（评分多），具有独特偏好的用户推荐效果差。**用户冷启动问题解决方案：**1.提供非个性化的推荐，直至收集足够多的个性化数据 2.借助用户的其他信息（如个人信息、其他网站记录） (存在侵犯用户隐私的问题) 3.诱导式推荐：选取少量具有代表性和多样性的物品进行推荐(通过迭代式收集用户反馈，快速获得近似用户画像)。**物品冷启动问题解决方法：**1、采用混合方法解决Item维度的冷启动问题，如引入基于内容的方法，通过物品的元信息得到初步推荐结果。2.一些边信息（Side Information）也有助于补充更多依据。如众包文本、评论信息、知识图谱。**基于模型的协同过滤：**基本思路：用户对物品的评分，本质上是用户的偏好，与物品的属性之间的相似度，相似度越高，评分越高。引入潜在因素，来表示用户的偏好与物品的属性。**基本形式：**借鉴矩阵分解的思路，揭示潜在因子。评分矩阵R(M\*N)被近似视作物品属性矩阵Q(M\*k)与用户偏好矩阵P(N\*k)转置的乘积。问题：**潜在因子的维度k如何确定？**当用户与物品的潜在因子已知，则任何缺失的评分，均可以通过对应的 P、Q矩阵相应的行列运算估计得到： 。**如何估计出两个潜在因子矩阵？**，是一个误差平方和（SSE），通过优化SSE，获得潜在因子的估计值。需要训练的参数一共(M+N)\*K个，随着K的增长而增长。如果K足够的大，在数据稀疏的情况下，可能训练样本不够->**过高的K可能导致过拟合**。**解决思路：**1.更多数据；2过拟合与正则化：引入正则项，避免过大的参数值。**衍生模型：非负矩阵分解**(优化目标里加入非负约束，最小化噪声矩阵E=R-PQ，利用梯度下降法优化，选取合适的学习率使得迭代时保持非负)；**概率矩阵分解：**利用参数本身规律(符合高斯分布)，假设1：参数与噪声均符合高斯分布；假设2：两个潜在因子矩阵相互独立，各用户/物品独立同分布。**矩阵分解扩展：社交约束**：好友在偏好与行为上十分相似；但：用户在不同社交关系中体现出不同偏好影响。

**Ch15社会网络（上）**主要讨论无向图，对称关系，如朋友、合作关系

**社会网络基本元素：**节点(实体)，边(关系)，超边/超图(点集间的边)。V的邻居集合N(v),出度和入度(一个网络里，节点出度之和等于入度之和等于边数；真实网络里节点度数符合幂律分布：少数节点拥有大多数边)。连通性：强连通(任意两点有双向路径)，弱连通(忽略方向时任意两点有路径)。连通组件(连通子图，子图任意两点间连通,同样区分强弱)。**节点的角色**：意见领袖：地位和作用近似HITS中的权威（对其他节点施加影响）和枢纽（为其他节点提供信息）。结构洞：主要作用在于为组织引入外部信息（判定方式：如果移除该节点会使网络变成多个连通组件）；结构洞的另一衡量方式：聚集系数（该点的任意两个好友也互为好友的概率）越低，该节点作为中介作用越大；意义：各方沟通的桥梁。**链接预测：**嵌入性：一条边两端节点共同的好友数。嵌入性越高，两端节点之间联系越紧密。**基本链接预测方法（CN）**：两点间有边的概率正比于共同好友个数，合理性为共同好友越多潜在交集越大，被介绍认识的机会多。**改进版（4种）**：削减好友个数的影响(防海王)，进行normalization。**无尺度网络（PA）**：各节点度数间差异较大且符合幂律分布，好友多的节点更易获得好友，在很多场景下符合实际，好处之一**在于无需网络结构信息和邻居信息**。**考虑邻居的价值（RA）**：不仅考虑共同好友，还考虑好友的好友数，共同好友越多，且共同好友的好友数越少则越好。**从局部到全局（Katz）**：不仅考虑共同好友，还考虑共同好友的好友，**基于多跳关系的链接预测基本方法** An表示沟通x、y节点的，路径长度为n的路径数量，N=1时，A即为共同好友数。同时，β为调节路径权重的参数，β 较小时，Katz指标其实很接近于共同邻居指标 。**一般化基于全局网络结构的链接预测方法**（Random Walk with Restart）：，其中， e\_x向量仅x节点对应的维度为1，其他为0，相当于指定起点，即预测目标的两端节点 ;q\_x为向量。RWR指标计算公式。**局部到全局+有监督：**Random walk（全局网络结构）+Pairwise LTR（训练偏好）。**网络表征：**将图转化为节点的序列，采用类似文本的表征方法。Node2vec，从BFS/DFS两种路径获取序列。**网络的符号性**：朋友/敌对关系，满足：朋友的朋友是朋友，敌人的敌人是朋友的关系三元组是平衡的。若图中所有三元组都平衡，则该图也是平衡的(一个图的节点可以被分为两组，组内的节点都是朋友关系，而组间的节点都是敌人关系，则该图稳定).可基于符号网络实现二阶约束。**链接的稳定匹配问题**。**社团挖掘基于层次聚类的社团挖掘：**分裂式聚类。本质：逐步去掉弱连接的过程。将沟通不同连通组件的桥接逐步删除。边介数：网络中任意两点间的最短路径，有多少条会通过这条边。Girvan-Newman算法：基于边的边介数实现节点层次聚类。1.计算网络中所有边的边介数；2. 去除边介数最高的边；3.重新计算去除边后的网络中所有边的边介数；4.跳至步骤2重新计算，直至网络中无边存在。局限性：无法确定合适的社团数（聚类方法通病）。选择社团数：1.采用层次聚类的评价指标进行选择，如社团的凝聚度、分离度；2.**Q-Modularity指标**：。选取Q-M值最大时的社团数。其中仅当ij节点属于同一社团时为1否则为0。A为网络对应邻接矩阵，k是结点的度，m为边的总数量。中括号部分，是节点之间的实际连边概率高于期待值的程度。 (ki / 2m)\*(kj / 2m) 表示如果节点度数不变，所有的边打乱重来，恰好在i和j之间仍存在一条边的概率。(Aij / 2m)表示当前情况下两节点之间存在边的实际概率。主要为了判断，相比随机情况，现有社团的内聚性有多好。**基于划分聚类的社团挖掘**：1. K均值聚类。节点属性：采用邻接矩阵，每个节点一个n维0/1向量,1表示邻居。2. 谱聚类：基于最小割算法，面向图结构的聚类技术。

**Ch16社会网络（下）主要讨论有向边，非对称关系，如通讯、关注关系**

**信息传播的元素**：发送者、接收者、媒介(传播过程发生的平台，如社交媒体)。信息传播的模型：依赖显性网络：本地信息(微信好友)/全局信息(大V)；依赖隐性网络。本章讨论的是显性网络的本地信息类型。**基本传播模型**信息级联：描述人们在信息流中的从众行为，信息级联过程描述人们传递他们从好友那里获取的消息的过程。信息级联基本假设：1.信息级联发生在一张有向图上（无向图转化为双向边），边是信息传递的媒介。2.每个节点仅能将信息传递给与其直接相连的节点。3.网络中节点状态是二元的。（激活：已收到并转发信息；未激活）4.已激活的节点才具备激活其他节点的能力。5.激活是不可逆的过程(激活后不能退回未激活)。**独立级联模型（ICM）**每次激活都是一次独立事件(抛硬币)，每个已激活节点只有一次机会尝试激活它的未激活邻居节点。轮次：如果某个节点在第 t 轮被激活，那么，他仅有一次机会，即仅能在t+1轮尝试激活他所有未被激活的邻居节点，节点v激活邻居节点w的概率为Pvw。整个传播过程直到所有节点都被激活，或没有新节点被激活为止。t=1时仅种子节点可尝试激活其他结点。**线性阈值模型 LTM**将信息传递过程视作多人影响的叠加过程。一个用户会被某个信息激活，如果来自他已激活邻居的影响超过某个阈值 。阈值预先设定，往往为从[0,1]均匀分布中随机抽取的一个数值（也可根据用户对信息的兴趣决定），阈值一旦设定，那么信息传播过程就一定了，不像ICM存在随机性。 仅有 t 时刻被激活的节点才能在 t+1 时刻施加影响。**传播最大化问题：**假定初始的种子节点集合为S，预期激活的节点集合为f(S) 。信息传播最大化的目的，在于限定S集合的前提下，最大化f(S)的规模。**常见的约束为限定S集合的规模，即|S|**。如果S集合中的节点价值不等，则可将约束进一步扩展。**启发式思路：**寻找网络中最具影响力的节点。启发式方法：1.pagerank及其衍生模型如HITS。2.核心性度量：用于衡量网络中最重要的节点。常见度量如度、紧密性、介数。3. 计算单个节点所能够激活的邻居数量，再进行排序。度核心性：思想:某个节点的邻居越多，对于网络连通和信息传递的贡献越大。背景事实:许多现实世界网络具有无尺度性，只有少数节点度比较高。归一化时的n为结点总数..紧密度核心性：思想:某个节点到其他节点越近，对网络连通和信息传递的贡献越大。背景事实:距离越近，信息到达所需的中介越少，传递越快，成功率越高,,g()为两点间最短距离。介数核心性：思想:越多最短路径经过某节点，对网络连通和信息传递的贡献越大。背景事实:与边介数思想类似，介数越高，信息传播中的中介地位越显著。，表示结点s与t间最短路径经过节点i的数量，上述启发式方法，在寻找“最具影响力的节点”时可行，然而，在确定影响力节点集合时不可行：节点的影响范围之间可能存在重叠。**一般化的信息传播最大化方法 ：**基于前述的传播模型(例如:ICM / LTM)，对于给定 S 集合计算 f(S)。此时，传播最大化问题可转化为一个带约束的最优化问题：找到S，使得f(S)期望（因为是随机过程）最大，且|S|=k。f(S)性质：1. f(S) 函数是非负的(显而易见) 。2. f(S) 函数是单调非减的，即f(S+v)≥ f(S)（新增加一个节点，至多不增加新激活，不至于减少）。3. f(S)函数是具有子模特性的，即: 对于任何集合对S，T，且满足S ⊆ T时，给定节点 v，有 .信息传播最大化问题解决方案：坏消息：在ICM / LTM等模型定义下，传播最大化问题是个NP难问题（简要证明思路:将这两个模型定义下的传播最大化问题，归约为集合覆盖和节点覆盖问题）好消息：由于 f(S) 函数具有子模特性，我们可以采用贪心算法近似求解。贪心算法：以空集合为起点，即初始 S = ∅。经过 k 次迭代，每次选择最大化 𝑓(𝑆 ∪ {𝑣}) − 𝑓(𝑆) 的节点 v。贪心算法可以实现至少 (1 – 1/e) 的近似效果，所得S可以激活至少 63% 最优解能激活的节点数。**衍生传播模型：独立级联模型的局限性**：1.每个节点只有一次传播信息的机会，是否过于苛刻？实际情况下，只要信息还在，就可以持续输出影响。2.节点状态未必二元化，也难以获得清晰明确的激活轮次,ICM 适合类似微博等具有明确转发记录的场景，但很多场景下并没有明确的信息传播轨迹：节点是否真的被激活？何时被激活？ 无法回答。**稳定状态传播模型(SSS)**：节点状态不再二元化，而是引入一个变量表示当前被激活的概率, 如果被激活概率不为0，则节点可以持续对外输出信息 / 影响。**SSS的隐患**：不合理的反向传播(因为结点只有激活概率)，方法1：遍历网络，找到所有类似这样的子节点(耗时费力，且结构特性难以判断).方法2：限制每个节点的迭代次数，超过一定轮次阈值，则该节点被激活的概率不再更新，阈值大致根据节点到信息源的最长 / 短路径决定。3.信息传播与接收的捆绑：接受!=转发。解决思路(实现传播与接收的解绑)：**1. 考虑信息覆盖问题**。核心假设：如果某个节点被激活，那么他的所有邻居都被覆盖，衍生出信息覆盖最大化问题->贪心算法。**2.修改模型框架**，对信息接受过程单独建模。引入函数 fu(Au)，描述 u 节点接受信息的概率，这一概率与有多少个邻居已经接受了信息正相关。进一步引出**信息接受最大化问题**：什么种子节点集合会导致接受信息的节点数期望最大，该问题在fu(Au)符合一定特性时，同样与信息传播最大化具有类似属性；也可将其视为更一般的框架，满足某些条件时，可退化为信息覆盖/信息传播最大化问题。

-------------------------------问答--------------------------------------------

**1.**主成分分析的基本流程是什么?与特征值有何关系?:

特征值对应的特征向量就是理想中想取得正确的坐标轴，而特征值就等于数据在旋转之后的坐标上对应维度上的方差。最大特征值对应的特征向量可以最大化投影方差。

**2.**基于机械分词的常见方法中对于“最大匹配”的依赖，可能导致什么隐患？如何利用 N-最短路径缓解这一隐患？如何选择一个恰当的 N 值？**最大匹配的隐患**：1.有长度限制：预设词长过短时长词会切错；过长时效率较低2.导致歧义： e.g. 南京市长春药店 南京市长江大桥。**N-最短路径**：基于Dijkstra算法，记录N条最短路径。原文中， N=2时， 非统计粗切模型句子召回率达到99. 73%； N=8时,句子粗切召回率达到了99.90%。**选择恰当的N值:**1.统计大量句子分词时最终选取是第n短路径，取这些n值的平均后可得到一个N值 2.或是通过机器学习、深度学习方法，构造包含N-最短路径算法的模型，利用人工标注了句子正确分词方式的训练集进行训练，确定N-最短路径算法中的参数N 3.或是考虑设置较大的N值来训练词的边权重：对找到的N条最短路径，按照正确切分方式人工对边权重做新标注。在实际使用时即可设置较小的N值来节省存储空间，减少人力复核消耗等，同时因为前面做过一轮边权重标注，所以又能保证一定的准确

度。

**3.**如何结合查询词项的分布细节，设计相对合理的跳表指针步长? 索引分布密集处使用较大的步长；索引分布稀疏的地方使用较短的步长

**4.** 在信息检索系统中，如何同时使用位置索引和停用词表？潜在问题有哪些，如何解决？

**如何同时使用**：一般方法是先对词典用停用词表过滤一遍，再在倒排表构建里使用位置索引，这样可以节省大量空间，因为停用词多为高频词，在大量文档中出现，提前过滤掉停用词可以让倒排表所需存储空间大大减小。对于已经利用位置索引建立好的倒排表，若还含有大量停用词，可以考虑用停用词表来对其过滤。**潜在问题：**某些停用词在特定场景下有意义，某些停用词的组合有意义，位置索引对存储空间需求较大，因为还要存位置信息。**解决方案：**进一步可以考虑，先用位置索引的方法构建倒排表，然后利用有实际意义的停用词组合构成的表来对倒排表作过滤，比如：对于停用词的组合：to be or not to be, 考虑对to 和be的倒排表作检查，利用其位置信息，若to 和 be在同一文档中的位置符合这样一种关系，则考虑不作去除。没有这样特殊关系的停用词则去除。当然，上述方式在初始构建倒排表时开销较大。所以也可以借鉴多元词索引的方法，把有意义的停用词组作为一个词汇加入词汇表，在构建倒排表的过程中，就可以把无实际意义的停用词去掉。在存储空间方面，考虑利用索引压缩技术，分别从词汇表和倒排表入手作压缩。具体到位置索引的使用上，可以考虑把存储位置信息也改成存储位置信息间的距离，并利用可变长度编码。

**5.如**何从多源情境信息（如手机的多种传感器信息）中，抽象出用户当前所处的状态或行 为模式？ b) 在上述过程中，如何既体现用户的个性化因素，又减少用户个人记录稀疏的负面影响？**答：**多源情境信息包括：陀螺仪、手机定位信息、手机噪声检测功能、摄像头、麦克风、温度传感器信息、用户手机APP使用记录等。由此可以想到，若能得到用户在某一段时间内的位置信息，如停留不同，再结合时间(如00:00)，可初步判断用户在家里休息。若还能检测到用户对摄像头、麦克风等的调用信息，还可以判断用户在进行使用手机聊天、视频的行为。如上所述，可以对用户一天、一周、一月内各个时间段持续监测，那么可以有效刻画出用户特定的行为模式。**第二问：**首先可以考虑采集广大用户(可能需有一定量的志愿者愿意提供准确数据)的一些数据构成训练集，并利用诸如用户身份(学生、白领)，用户手机APP种类等对行为模式具有导向性的特征对用户群进行划分，利用训练集对每一类用户的行为模式做一个笼统的模型描述。体现个性化因素：考虑将用户使用app的种类和时长，以及用户各时间段的位置信息等纳入考量，为用户在基础模型上进行个性化微调。减少稀疏记录的负面影响：因为已经有一个基础模型，所以当某用户的有关数据较少时，可利用同类用户的行为信息，结合基本模型对用户的行为模式有较好的刻画

**6.**用户在浏览网页时，可能通过点击“后退”按钮回到上一次浏览的页面。用户的这种回退行为（包括连续回退行为）能否用马尔科夫链进行建模？为什么？ **答：**不能。马尔科夫链的下一状态只与当前状态有关，回退行为需要记录之前的状态，两者矛盾

**7.**如何在网页排序的同时提升结果的多样化水平？如何在实现这一目的的同时保障算法的效率？修改优化目标，直接在将多样化指标写入目标函数扩大召回率，增加精排、重排模块 （**1**.在排序算法中使用多样性有关的评价指标，将多样性纳入排序结果的优化任务中。在获得初始的相关性排序结果基础上，对相似度高的结果做一定惩罚，使得前N个排序结果保持一定的差异性，从而得到具有多样性的排序结果。评价指标即考虑是用户查询与文档的相关度，文档间差异度两方面的一个加权考量。通过此方式迭代，可获得⼀个既有相关性也有多样性的检索集合。**2.**可以考虑引入主题划分，不同网页有各自的主题归属。这样不必对排序后的所有网页判断相似度，因为让所有网页都判断相似度计算量太大。所以只是考虑取各个主题前k个结果(或根据各个主题与用户意图的接近程度分配不同权重)，最终能保证获得的结果中包含有不同的主题，同时保障了算法的效率）

**8.** 如果从信息检索的视角， 可以将寻找最近邻的过程视作检索最相关的 K 个文档的过程。那么， 这一过

程是否可以利用倒排索引的思路加以实现？如何实现？**可以**。回顾第6章讲过的向量空间模型（VSM），可以把文档与查询表示成词项的tf-idf权重向量，再计算相似度，按相似度大小排序后返回Top K的文档。**如何实现**：引入倒排索引的思路，考虑查找样本的k个邻居时，只查找与待分类文本的词条有重叠的文本（由此减少搜索空间）1. 倒排索引结构包括词条数组和每个词条的文本链表，词条数组是将所有训练文本分词、特征提取后的所有特征项组成的数组，每个特征项为一个词条。每个词条的文本链表则包含含有该词条的所有文档，每一项表示为文本ID+词条在该文档中的权重(如tf-idf)。并且链表是按权重值排序过的。2. 把待分类的查询文本d表示为文档向量V，找到向量V中每个词条的文本链表，将这些链表合并，去除重复文本id的项，最后计算查询与这些文档的距离，返回Top-K（k最近邻）再进一步，考虑通过聚类把训练文档划分为K个区域（簇），每个区域用一个中心向量表示，并建立词条到簇的倒排索引。利用倒排索引找到与查询文本有交集的簇，先计算查询文本与簇中心向量的相似度，超过阈值的簇保留，并让查询文本与簇内每个文档计算相似度，最后返回Top-K。

**9.** 无论是 K 最近邻分类还是 K 均值聚类，都涉及到 K 的取值问题。请简述两个问题各自选取合适 K 值的思

路，并比较两者在思路上有何不同？K最近邻分类：在训练集上， 使用不同的K进行分类， 选择分类效果最好的K。K均值聚类：尝试使用不同的K值聚类， 检验各自得到聚类结果的质量， 选择

聚类效果最优的K。基本思路本质上是一致的

**10.** **阐述K-mediods算法和K-means算法相同的缺陷：**必须事先确定类簇数和中心点， 簇数和中心点的选择对结果影响很大；一般在获得一个局部最优的解后就停止了；对于除数值型以外的数据不适合；只适用于聚类结果为凸形的数据集等**。阐述K-mediods算法相比于K-means算法的优势：**与K-means相比， K-mediods算法对于噪声不那么敏感， 这样对于离群点和异常点就不会造成划分的结果偏差过大， 异常数据不会造成重大影响。**阐述K-mediods算法相比于K-means算法的不足：**kmediod需要不断的找出每个点到其他所有点的距离的最小值来修正聚类中心，这要求更高的计算复杂度， 减缓了收敛的速度， 因而对于大规模数据的聚类力不从心。

**11.** 社团推荐是社会网络分析中常见的问题。在仅考虑网络结构的情况下，已知网络中存在的若干社

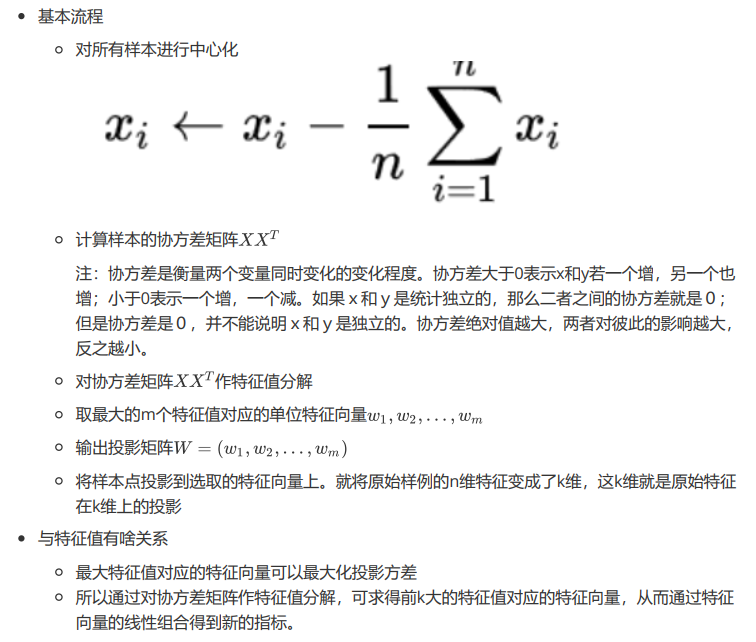
团，请结合课程内容中所介绍的推荐技术（包括链接推荐技术）与社团挖掘技术，设计一种给指定节点

推荐社团的方法，并简述你的理由。**答：**先利用链接预测技术（如RWR指标）为指定结点补全其在原网络结构中的关系。**理由：**充分挖掘指定节点在网络中的信息与关系，为下一步社团挖掘提供更多依据。

再将各节点利用层次聚类（如Girvan-Newman算法）划分到某个社团，这里实际上因为已知社团

数，所以划分过程比较直观简单。当然也可根据Q-Moduality评价指标来筛选，比如尝试多种链接预测技术（基于多跳关系/有监督学习/仅考虑共同好友数量以及共同好友好友数等），来补全网络，然后比较Q-M值来作优选，使得社团内聚性更强。

-------------------------------计算例子--------------------------------------



图片包含 设备

描述已自动生成可变长度编码：214577的二进制为 1101/0001100/0110001，编码为00001101/00001100/10110001。

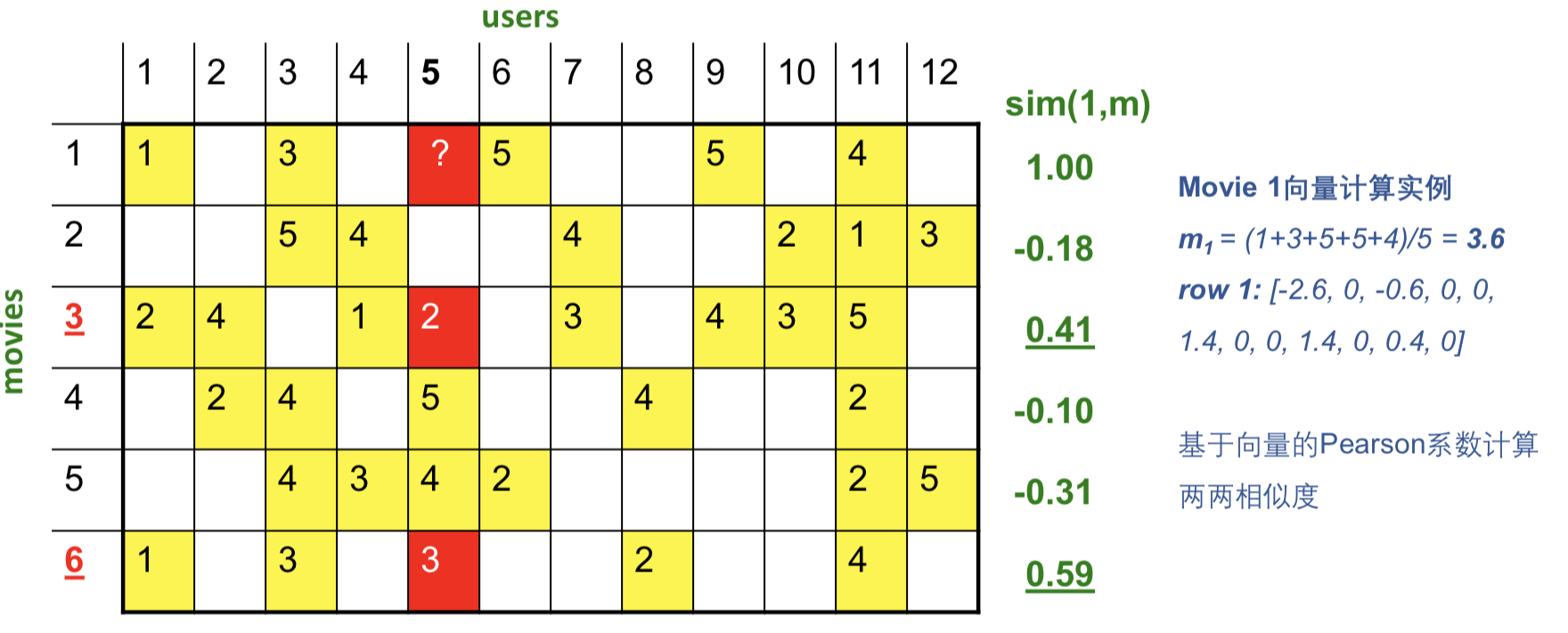
信息增益和基尼指数：（注：信息增益越大越好，基尼指数越小越好）

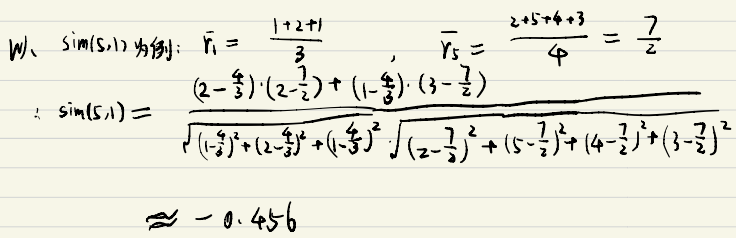
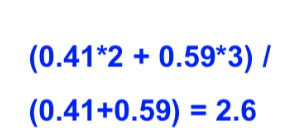
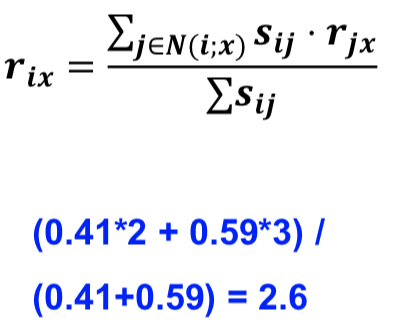
(a) 以属性A划分时的信息增益。

(b) 以属性A划分时的Gini指数。

协同过滤

基于物品推荐





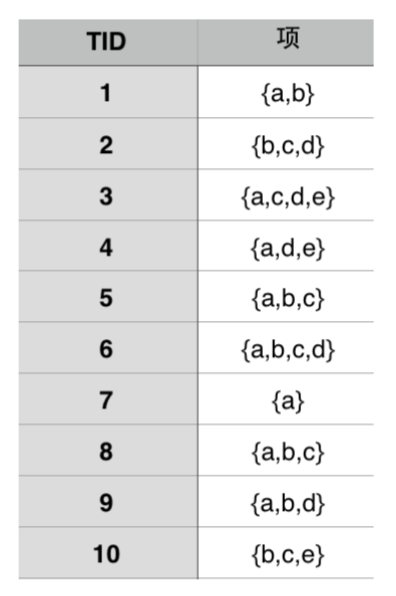
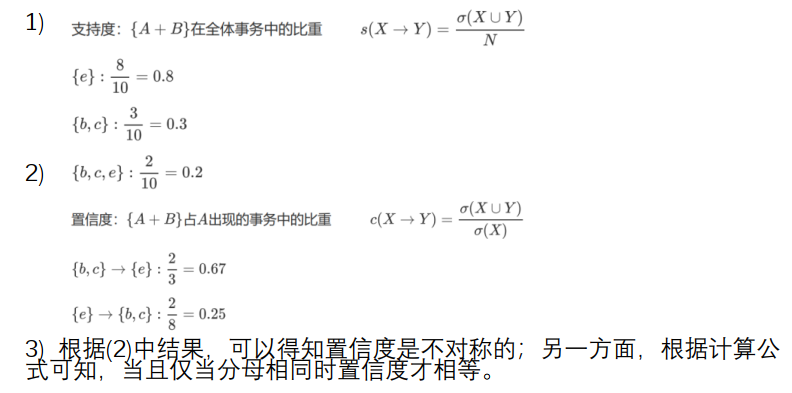
基于用户过滤:

基于用户推荐计算得到最相似的两个用户为3和9，相 似度分别为0.21和0.47 .

图片包含 物体

描述已自动生成

FP-Growth

图片包含 眼镜

描述已自动生成

