信息存储与检索历年考题（XWQ小弱版）

1. **信息检索的基本原理** P3

**我们可以把信息检索的基本原理抽象、概括为一句话:对信息资源集合与信息需求集合的匹配和选择**。

**信息资源集合**是指有关某一领域、经选择性采集和组织加工的信息集合体。信息资源集合是一种公共知识结构，它有可能弥补某个特定用户的知识结构缺陷，即可以向用户提供所需要的信息、知识或获取知识的线索，或者提供某种信息去激活人脑中存储的知识。

**用户的信息需求**是在社会实践活动中产生的。当人们为完成某一任务或工作时，经常会觉得缺少某些知识，这就产生了信息需求。众多用户不同形态的信息需求的汇集，造就了信息需求集合的存在。用户信息需求的产生及其满足，是实施信息检索行为的前提与基础，更是实施信息检索行为的目的所在。

面对信息资源集合与信息需求集合，如何在两者之间建立起联系与沟通的桥梁，以便能从信息资源集合中快速获取用户所需要的信息与知识呢？这就要求信息检索提供一种**“匹配”机制**。这种机制的主要功能在于：能够把信息需求集合于信息资源集合依据某种相似性标准进行比较与判断，进而选择出符合用户需要的信息。

1. **信息检索研究的主要问题？为什么说它是一个交叉领域众多的学科**P5

概括来说，信息检索关注的研究问题涉及理论、技术和应用等不同层面，主要有：

信息检索理论：**检索语言、检索模型、标引理论、相关性理论、知识组合与表示理论**、等等。

**信息检索工具/系统**：是由有序化的信息资源、设备、检索技术和检索方法等组成的有机整体，构成了实现信息检索活动的物质基础，是信息检索的现实研究对象。

信息资源及其收集、加工：主要涉及“信息存储”问题，其目的是建立和形成可供检索服务的各类机读数据库。

**检索技术与方法**：从检索对象的不同表现媒体上来区分，检索技术分别有针对文本、数值、音频、视频等信息的不同检索技术与方法。

**用户研究与检索策略**：用户是信息检索和信息检索工具/系统创建与生存的基础，无论是系统的研制开发、管理维护、功能和服务的扩展，还是系统评价等，都离不开用户研究工作。

其他密切相关的自动化处理技术：具体涉及到自动聚类与分类、自动摘要、信息可视化、信息过滤、信息提取、机器翻译、人机交互等众多方面。

**相关学科及领域**：从前面对主要研究问题的分析可以看出，信息检索是一个比较典型的交叉研究领域。虽然从起源上观察，图书馆学，情报学是信息检索的母体学科，他们对于早期信息检索活动的产生、发展、手工检索服务等提供了比较完善的理论知道和良好的应用舞台，但在目前的研究活动中，信息检索已更广泛借鉴、使用了来自其他自然科学、社会科学、人文科学等领域的研究成果与研究方法。与信息检索关系最为密切的相光学科及领域主要有：计算机科学，数学，系统科学，语言学与计算语言学，认知心理学。

1. **如何形式化表现信息系统？形式化的意义。P17**
2. **信息检索系统的基本结构**P53

系统的结构是指系统的组成及各组成部件之间的关系。系统的结构具有多维性，对于一个信息检索系统来说，我们可以从不同的角度和层次来讨论其基本结构。

检索系统的各种不同的数学模型可以看作检索系统的**概念结构**。

**物理结构**：**硬件部分**：系统采用的各种硬件设备的总称

**软件部分**：包括系统软件和应用软件

**信息资源集合**：目前以数据库的形式存在于检索系统中，除数据库之外，某 些只能检索系统还可能包括各种模型库、规则库、知识库等。

按照物理空间分布情况，信息检索系统的物理结构又可分为集中式和分布式两大类。

**集中式检索系统**：资源在空间上集中配置的系统。

**分布式检索系统**：是指通过计算机网络把分布在不同地点的计算机硬件、软件、数据库等设备和资源联系在一起，以服务于一个共同的系统目标而实现的相互通信，互操作和资源共享的系统。

**逻辑结构**：主要是指它所包括的功能模块（或子系统）及其相互关系。一个通用的计算机化的信息检索系统，通常应该具有“信息存储”和信息查询两大基本功能。

1. **论述一下三大经典检索模型** P16脚注

**布尔模型**：一种简单的检索模型，它建立在经典集合论和布尔代数的基础上。鉴于集合论中“集合”概念的直观性以及布尔表达式所具有的准确语义，布尔模型非常容易被用户理解和接受，在早期的大多数商业化书目检索系统中，布尔模型更是得到广泛应用。两个基本规则：每一个检索词在一篇文档里只有两种状态，出现或不出现。提问式由布尔逻辑运算符连接索引词构成。

**向量空间模型**：文本向量的构造，对于任一文档，我们都可以把它标示为如下t维向量的形式 提问向量的构造，在向量空间模型中，用户的信息需求被加工，转换为提问向量，并用与文档向量累死的表示形式表示 匹配函数的选择及相似度阈值的确定：在文档与提问向量化表示的基础之上，文档与查询提问之间的相关程度（即相似度）就可以由它们各自向量在t-维空间的相对位置来决定。

**经典概率模型**：给定一个用户提问，则检索系统中存在着一个与该提问相关的理想命中结果集合，这里我们不妨用R来表示之。如果能已知集合R的主要特征及其描述，则用户的检索要求不难实验。但，问题是，在用户提出检索要求时，并不知道这个理想结果集合的特性。为此，需要在检索伊始就对R的特性进行某种猜测。根据初始的猜测，系统将检索到一个初步的命中结果集合。由用户或通过相关性自主判断，不断优化和改进，逐渐接近理想集合。

1. **向量空间模型的原理以及特征**p26

**向量空间模型的基本原理**：

**文本向量的构造**，对于任一文档，我们都可以把它标示为如下t维向量的形式

**提问向量的构造**，在向量空间模型中，用户的信息需求被加工，转换为提问向量，并用与文档向量累死的表示形式表示

**匹配函数的选择及相似度阈值的确定**：在文档与提问向量化表示的基础之上，文档与查询提问之间的相关程度（即相似度）就可以由它们各自向量在t-维空间的相对位置来决定。

**向量空间模型的技术特征分析**

采用部分匹配策略，使得在算法层面上基于多值相关性的判断处理得以实现

采用基于统计学方法的词加权处理模式，使检索结果得到显著改善

采用对检索结果排序输出的策略，使对检索结果数量的控制与调整具有相当大的弹性与自由度。

**缺陷**：索引词两两正交假设

1. **阐述tf-idf法及其合理性 p27/P61**

**定义：**假设N为系统文档总数；ni为系统中含有索引词ki的文档数freqij为索引词ki在文档dj中出现的次数；idfi表示索引词ki的逆文档频率；maxtfj表示文档dj中所有索引词出现次数的最大值。那么对于文档d中索引词k的权值计算方法为：fij=freqij/maxtfj (局部权值) idfj=log(N/n)（全局变量） w=fij\*idfj（索引词权值）最后一个公式被成为“tf-idf”加权模式。

**合理性未找到。**

1. **信息检索系统的用户接口模块有哪几个功能模块组成？p67**

一个信息检索系统的用户接口模块通常是由用户模型、信息显示、交互语言和反馈机制等部分构成的。

**用户模型**：是指由检索系统设计人员建立的用户认知模型。（考虑不同用户的知识水平，技能和经验及其不同的信息需求状态等因素）

**信息显示**：是指检索系统以屏幕显示形式提供给用户的各种操作信息

**交互语言**：主要是指系统提供给用户使用的检索命令集合和其他对话工具。

**反馈机制**：即检索系统对用户操作及其反馈的信息所做出的一种反应机制。

1. **视频信息的基本构成和检索特征**p151

视频是另一类重要的视觉信息源，它不仅包含静态图像所包含的内容，还包括场景中目标运动的信息和客观世界随时间变化的信息。

一个视频序列或视频节目讲述了一个比较完整的故事，它由一系列的情节构成，一个情节描述了一个具体的事件或行动，包括多个内容上相关的镜头，一个镜头对应了在一个空间场景拍摄的连续动作，用一组连续的画面来表现。

**视频索引**：

**基于视觉特征的索引**：适合各类不同的视频数据，因为视频特征是比较底层的检索特征，其特点是与领域相关性较小。在当前的技术条件下，视觉特征大都可以采用图像处理和分析技术比较自动地从视频中提取。

**基于语义特征的索引**：主要是对围绕目标的一些语义基元进行索引。获取视频语义特征的主要方法和手段是视频摘要和视频注释。视频摘要是指原始的视频序列进行加工以获得该序列的一个较为简明、紧凑的表达。视频注释是指利用一些描述性的手段，例如文字，图标，绘图等来描述视频的内容和语义。

1. **什么是自动标引？简述自动标引的流程和核心步骤。**P60

标引是指对信息资源的各种检索特征进行分析并使之显性化，以便为存储和检索这两个环节提供某种连接的一种重要的信息加工及操作。**自动标引是利用计算机实现对信息资源的自动标引处理。**

**流程：待标引机读文本——语句分析（抽词/切词处理）——候选标引词——词频统计——加权计算——阈值确定——标引词确定——（概念转换——规范词语选取——）索引文档生成**

**核心步骤：词语加权计算 主要方法：绝对词频法 逆文档频率法 词区分值法**

1. **中文文本自动标引的词语切分方法** p62

**词典切分法**：主题词表法，关键词词典法，部件词典法

**切分标记法**：用计算机扫描输入的汉字串，逐字与“非用字后缀表”进行比较，如果是“非用字”，则舍去；如果是“用字”，则取出。 依据预先拟定的不同构词模式，对抽出的字符串进行分解，形成单词或专用词组。 对抽取出来的单词或词组进行优化，形成最终的标引用词。

**单汉字法**：将中文自动标引的基本单位由“词”一级将为“字”一级，其标引处理的基本思想是:从文本中将汉字注意去除，同时记录下它们的文献号、字段号以及汉字所处的位置编号，然后把这些信息写入每个汉字建立的索引文档中。

**N-gram法**：与语种无关，指由N个相邻字符组成的字符串序列。

1. **论述一下基于内容的检索系统 p16**
2. **用一种提问式的变换形式说明布尔检索的技术实现P88**

**A+B\*(C+D) +A\*B+CD ABCD+\*+**

1. **信息检索的技术和辅助检索方法**
2. **一个文献检索系统的结果后处理软件工具应该具有哪些功能？p117**

排序输出目前大多数检索系统都采取了某种检索结果排序输出的处理方案，最为常用的排序标准是相关度排序，即按照检索结果条目和用户检索提问式之间匹配程度的大小从高到低顺序输出结果联机聚类与自动摘要 对庞大检索结果集合的另一优化处理方案是对结果集合进行某种可视化方式的组织和显示，以便用户比较只管、快速地从中筛选出最想得到的信息去重合并 检索结果中存在重复的记录信息，目前独立搜索引擎对检索结果中的重复信息很少进行过滤处理。不过，一些比较优秀的元搜索引擎系统则普遍具有对来自不同成员引擎重复检索结果的合并和去重功能。

1. **图像的基于内容检索类型，结合实例回答**p138

**基于颜色特征的图像检索**：颜色直方图法，颜色对特征分析法，主色调分析法

**基于纹理特征的图像检索**：使用文理测度的方法和使用感知属性的方法Netra系统和CANDID系统 QBIC系统

**基于形状特征的图像检索**：特征方法，形状变换法，关系方法

**基于空间关系的图像检索**：使用关键词检索，使用范图查询，使用草图查询

1. **元搜索引擎的工作原理**p169/p183

在使用元搜索引擎时，用户只需提交一次检索请求，经转换处理后，检索请求可同时转交给多个预先指定的独立搜索引擎去查询，然后将所有查询结果汇总起来再以统一的格式呈现到用户桌面上。

由于串行元搜索引擎只是给用户提供一个独立搜索引擎列表，由用户直接选择所需要的引擎并加以调用，所以在实际应用中，元搜索引擎主要以并行工作方式为主。

基本结构：用户接口、查询代理和结果汇总输出。用户接口是用户浏览器与元搜索引擎交互的界面。其主要任务是接收用户的查询请求，完成用户个性化的检索要求设置，并进行查询请求的转换和分发。查询代理负责元搜索引擎与相应的独立搜索引擎的交互。对于并行工作的元搜索引擎来说，它链接的每一个成员引擎都对应着一个查询代理。每个查询代理根据接收到的用户查询请求调用其代理的成员引擎进行检索，然后，各成员引擎将所得到的检索结果分别返回各自的查询代理。结果汇总输出负责收集各查询代理送交来的检索结果，并对这些结果进行必要的加工和整理，再以标准页面的形式呈现给用户。

1. **独立搜索引擎的工作原理和目前遇到的困**境p163

独立搜索引擎的基本结构一般应该包括数据采集、数据分析与标引、数据检索、信息挖掘等几个不同的功能模块。

**数据采集**：目前大多数的搜索引擎都是采用自动方式进行数据采集。

**数据分析与标引**：同样分为2种。人工方式主要用于目录浏览型搜索引擎。自动方式是通过自动标引软件对收集到的网页信息进行内容分析，从中提取到有检索价值的特征项——网页关键词等，并将它们组织成倒排文档。

数据检索：主要负责通过用户界面的人机交互，根据用户的信息查询请求，或者去索引数据库进行查询匹配，或者利用分类目录结构进行浏览选择，最终将获取的符合用户需要的WWW信息显示输出。

数据挖掘：跟踪、发现用户的需求兴趣，建立用户兴趣模版文件，并利用兴趣模版文件过滤检索结果，以提高检索服务的质量。

**目前的问题**：Index 1，全文标引还是部分标引2，是否过滤停用词汇3，是否使用Meta标记的信息4，是否支持词干提取技术5，是否对图像标记中的替换文本或者页面中的注解建立索引

1. **相关性的定义、性质、评判标准p193**

简单地说，信息检索的“相关性”主要是指检索系统针对用户的信息需求从文档集合中检出的文档与用户需求之间的一种匹配关系。

**性质**：关系是“相关性”最核心的。 直觉的，对任何学科来讲，要给某一直觉概念下精确定义都是很困难的，而且无论何种定义都会存在商榷的余地。 多维的，“相关性”是一个多维的认知概念。动态的，相关性的动态特性是非常明显的。

**判断标准**：三种。二值相关性判断标准 相关文献1 不相关文献 0 多值相关性判断标准 多维相关性判断标准 （信息资源 用户需求 时间 组件）

1. **等级聚类法的原理和性能**p226

等级聚类法是文本聚类处理中应用较多的一类方法，它通过建立并逐步更新距离系数矩阵（或相似系数矩阵），找出并合并最接近的两类，直到全部聚类对象被合并为一类为止。

**步骤**：计算文档的距离系数矩阵 合并两个最相似的文档类 更新相似矩阵 重复2和3 直到合并成一个类

**特性**：每一聚类步骤都是将两个文档（或文档类）聚集成一个新类，因此，全部聚类过程需要n-1次循环 他们是“贪心”的，即每次聚类总是选择在当时情况下最相似的两个类进行合并 它们是全局的，在聚类过程中所有的类间关系（相似度）都被考虑到了，因此在给定了聚类结束条件的情况下，其聚类结果是稳定的。 聚类的时间复杂度是O（n2），其中n是文档数。

1. **自然语言的研究，包括文本检索在内的一系列文本处理自动化技术主要形成奠定了哪些共有的基础性研究思路或者策略。P222？**
2. **什么是自动文摘？简述自动文摘的技术，并评价p256**

文摘是指对一份文献（或文献单元）的内容所做的简略、准确的描述，并且通常不包含对原文的补充、解释或评论。**自动文摘即利用计算机实现文摘编写与生成的自动化，它是计算机技术、自然语言理解技术和传统文摘工作相结合的产物。**

**自动文摘的步骤**：**文本分析**：主要用于寻找最能代表原文内容的成分

**文本转换**：需要对文本分析产生的结果及其表示进行修剪和压缩，以便形成文摘

**文摘生成**：根据用户要求具体决定

**自动文摘的类型**：通用文摘和偏重文摘；单文档文摘和多文档文摘；有监督学习文摘和无监督学习文摘；

**基于统计的自动文摘**：适用面广，不需要复杂的语言学的知识，处理速度快，便于移植

自动摘录的质量不稳定，不全面，不简洁，不连贯

**基于理解的自动文摘**：牺牲领域宽度来换取对文章的理解深度，质量较高，实用水平还比较低

**基于信息抽取的自动文摘**：领域受限，仅凭文本特征词或特征短语的提示作用来填充文摘框架，有时并不能做到非常准确，实用文摘模版生成文摘，使得文摘的语言千篇一律。文摘效率和灵活性较高

**基于结构的自动文摘：**既能避免自动摘录的不连贯性，又能避免基于理解方法和基于信息提取方法受专业限制的缺陷。不过，基于结构的自动文摘法因为不能做到让计算机真正理解文献的主题内容，仍然具有很大的局限性。

1. **动态聚类 p229**

**定义**：动态聚类法又称“逐步聚类法”，它主要致力于在一个平台层次上分割所有的样本点，并通过算法的迭代执行，得到一个较合理的、有k个类的聚类结果。

**基本算法思想**：确定聚类个数k，从文档集合中选择最初的k个文档作为凝聚点，每个凝聚点文档自成一类按照距离最近原则，将剩余(n-k)个文档逐个并入最近凝聚点所代表的类。每并入一篇文档，立即计算该类的重心，并用此重心替代原来的凝聚点以最后形成的每个凝聚点代表一类，将全部n篇文档重新聚类，逐个并入最近的凝聚点所属的类。与步骤2相同，每并入一个文档后，就重新计算重心，并以此重心代替愿凝聚点。文档集合被重新聚类后，如果与原来的聚类结果不同，就重复步骤3；否则，聚类处理即告完成。

**两个关键问题**：如何确定并调整聚类参数k，如何选取合适的初始凝聚中心。

1. **介绍一下TREC(定义、目标、学术贡献等)p192/p212**

**定义**：1992年，由美国国家标准与技术局和国防部高级研究计划局共同发起并主板了“文本检索会议”。TREC是一项致力于对文本信息检索技术进行大规模评价研究的试验活动。

**目标**：多个目标（5个）促进基于大型文献集合的检索研究。通过提供大型的语料库和实验数据，来满足和应对当前信息资源因爆炸性增长而形成的海量规模状态，使检索系统的设计和算法更加符合现实生存环境的要求。建立一个开放的论坛来交流研究思想，以促进企业、学术机构和政府部门之间的合作、交流和沟通。通过展示新型检索技术在解决实际问题中的有效性来加速实验室项目向商品化产品的转换。如果某种先进技术表现出良好的实验效果，那么它就有可能比较迅速地得到商业化开发，从而达到技术先进性与解决实际问题有效性的有机结合方便企业和学术机构得到适当的检索评价技术以及开发更适合于当前检索系统的新的评价技术引导、开创信息检索领域新的重要研究分支和研究方向，丰富信息检索的理论、技术与方法研究。

**学术贡献**：TREC是成功的。作为促进和推动信息检索技术发展的有效举措和交流形式，它首先为理论检索模型和试验检索系统提供了公平、定量、具有实用价值的性能评价机会，并对评价结果位于前列的高性能检索系统提供了商业化的机会。

**组织形式**：TREC检索评价活动的组织形式可以概括为：年度周期、春季发布试验数据、夏季展开检索实验、年终会议提交评价结果。