信息检索系统

信息检索系统及类型

信息检索系统是不是信息管理系统/信息系统?
解释信息检索系统的含义

信息检索系统的概念

系统(System)：由若干个具有独立功能的元素(或部件)组成的一个有机整体，这些元素(或部件)之间相互关联、相互制约，共同完成某些规定的任务与目标。
特性：整体性，关联性，层次性，目的性，适应性。
信息检索系统（Information Retrieval System）：一类具有信息存储和检索功能，面向一定用户的信息服务设施。

基本要素（6个）

目标——明确的服务对象、专业范围、用途；（服务用户）
功能——检索及其它信息服务功能；（核心功能是检索）
资源——各种类型的信息，经加工后的有序集合；
设备——存储信息的载体、匹配选择的设施、输入/输出/显示/传递（通信）等设备；
方法——一定的处理方法，实现检索系统的功能；
人员——系统人员和系统用户。
在信息时代，信息检索系统通常是一类基于计算机及网络的人机交互检索的系统。

信息检索系统的类型

按设备划分

按信息存储与检索所用的设备
①书本式检索系统
以传统书刊形式提供的一类出版物，如文摘杂志、题录、书本式目录、参考工具书等。
②卡片式检索系统
图书馆内的各种卡片目录，管理部门的各种卡片档案。
卡片：记录，字段（表中）
③缩微式检索系统
以胶卷、胶片为信息存储介质的检索系统。通过检出和显示缩微画面供用户阅读。
包括：缩微胶卷检索系统、缩微平片检索系统、计算机辅助缩微品检索系统。
④计算机情报检索系统
使用磁带、磁盘作为存储介质，采用计算机控制存取大量信息。高级形式是联机检索系统。
联机检索：由大型计算机系统、数据库、一批检索终端等来实现，采用分时处理方式。
⑤光盘检索系统

以大容量的光盘存储器为数据库介质，利用光盘驱动器和微机控制读取和检索信息。
特征：数据库和检索软件一并提供给用户；
数据更新周期为月、季度或半年不等。
⑥多媒体检索系统
建立在多媒体数据库基础上的检索系统。不同性质的信息存储于不同媒体上，进行一体化管理。
多媒体数据库：Multimedia databases，是把文字、数值、图像、视频、声音等性质不同的信息存储在不同媒体上，进行一体化处理和管理的新型数据库。

按功能划分

数据库有哪些?
信息系统有哪些?
管理信息系统，数据流程图，数据据点

许多信息系统都含有检索（查找）的机制和过程，存有数据文件或信息集合，能够响应用户需求并输出答案。因此均可视为广义的信息检索系统。
萨尔顿认为，以计算机为基础的广义信息检索系统，主要包括五种类型。

文献检索系统（DRS）

Document retrieval system，是早期的情报检索系统的主要体现形式，也是狭义的信息检索系统。
特征：文献信息数据量大，属性繁多，连贯性强→采用非结构化的文件结构。（图书馆多用）
每篇文献对应一个记录，记录中含有各种文献特征信息（书目数据，如主题词、分类号、作者、标题等）。
检索：获得书目引文和原文出处。
系统主要面向：文献资源的管理者和利用者；
主要用户：科研和教学人员。

数据库管理系统**（DBMS）

DBMS：为任何数据库的建立、操作和维护而设计的计算机软件，可用于数据、事实的存储、检索与维护。
主要面向：结构化数据查询和基层事务处理。
主要用户：事务管理员、行政职员。
目前的DBMS已经商品化，一般具有良好的通用性和可移植性。

自动问-答系统（QAS）

Question-Answering System，也称事实检索系统。
QAS是一种能直接回答人们提出的具体问题的计算机系统。目前基本上已经归如专家系统（ES）。
该系统主要由知识库和推理机制组成。

此类系统允许自然语言提问，并以自然语言回答。因此又需要解决自然语言理解的问题。
主要用途：支持专家决策，为知识的采集、存储、利用服务。

管理信息系统（MIS）

Management Information Systems，主要面向企业管理人员，对用于管理的信息进行采集、加工、分析、传输和保存的一种人-机系统。
MIS是以管理科学、数学工具和信息技术为基础。
应用场合：统计会计、生产管理、市场营销，以至各个领域。主要面向中层管理人员。

决策支持系统（DSS）

Decision Support Systm，是管理信息系统（MIS）的高级形式，主要面向中高层管理和执行决策提供辅助，为决策者提供了一种决策环境。
DSS通常由数据采集、数据库管理、模型库管理、用户界面等模块构成。输出的是建议、最优方案，或者不同方案的比较与排队。

分析、比较上述五类信息检索系统

共性——均含有信息存储和检索匹配的机制和过程；
差异——各自的内部结构、功能特征有区别。

ssci，美国科学信息研究所，社会科学引文索引

三大科技文献检索工具

sci，科学引文索引
ei，工程索引
istp，科技会议录索引

信息检索系统的基本结构

1.什么是计算机应用系统的硬件和软件
2.解释信息检索系统的结构

物理结构

硬件、软件、数据资源。

SQL，结构性查询语言
structured query language
索引文件 main项和交叉引用的名称标记索引条目生成索引

硬件部分

主计算机：CPU、内存。要能够适应技术系统的要求，速度要足够快；内存要适应海量信息处理。对联机系统，还需具备多道、分时能力。

输入/输出/传输等设备； 检索终端；
外部存储器，如磁带、磁盘、光盘等。

软件部分

系统软件和应用软件。
系统软件：用于计算机的管理、运行、维护（配置）。如：OS、语言处理程序，例行服务程序。
应用软件：为解决应用问题所编制的程序。如：数据采集、标引、建库、词表管理、检索匹配等模块。

数据资源

目前主要以数据库（DB）形式存在于检索系统中。DB是计算机信息检索的基础，是现代信息资源的体现。
检索系统的数据库有多种类型：

参考型数据库

（reference databases），也称指示性数据库。包括书目数据库（属于二次文献）、指南数据库（如产品数据库）等。

源数据库

(source databases)。提供原始资料或具体数据。包含数值数据库、文本数据库、全文数据库、术语数据库等。

混合型数据库(mixed databases)。可存放多种不同类型数据。

逻辑结构

构成信息检索系统的功能模块或子系统及其相互关系。

信息组织：序化
标引
数据库：管理数据，操作，存储

一个完整的检索系统通常由6个功能模块构成：
信息源选择与采集子系统、标引子系统、数据库子系统、词表管理子系统、提问处理与匹配子系统、用户界面子系统。

信息源选择与采集子系统

（有的也称采选子系统）
功能：按系统既定方针和用户需求，从外部信息源选择并向系统输入。选择标准含专业覆盖面（全文信息、数值信息、书目信息）、文种、时间跨度等。
信息源——检索系统的信息或数据的来源。

信息源主要来自三种：
一次文献：作者以自己的研究成果为基础创作或撰写的、未经加工的原始文献。如，原始的期刊、图书、研究报告、学位论文等，构成全文DB
二次文献：对一次文献信息加工、整理而成的文献。具有汇集性、工具性、综合性、系统性。如，文摘、索引、目录，构成书目DB
三次文献：对一次、二次文献的综合、分析等深加工的产物。如，百科全书、词典、指南、年鉴、手册等，构成指示性DB**
基本要求：快速、经济、适用。

标引子系统

标引(Indexing)：借助一定的词表，对信息资源的各种检索特征进行分析，赋予标引词，形成标识，以便为存储和检索提供某种连接。标引以词表为基础。
主要任务：对信息资源中有价值的特征信息，如题名、分类号、主题内容、语种等进行提取与标识，为用户的查询提供准确而有效的检索入口。 （例如，对一篇科技文献的标引）
检索入口：又称检索点或检索标识，是指用以标识信息的外部特征和内部特征的属性值。如主题词、分类号、作者、标题、机构、代码等。它由人工或计算机生成，是用户检索的出发点或依据。
标引过程的基本要求：全面、准确、简洁，即不漏标、不错标、不滥标。
从技术上讲，标引可分为人工标引和机器标引。
人工标引：由标引员对文献或其它信息资源进行概念分析基础上实现的标引。
机器标引：由计算机对词的出现频率、出现位置、提问频率等进行统计与加权，选取索引词。又分为全自动标引、半自动标引等。
自动标引是一项与语种有关的处理技术。
如对英文，因词间有分隔符，利用计算机进行抽词和词频统计就非常容易实现。但对中文，词语间既无空格，也无任何特殊的间隔标志，故其自动标引要比英文困难得多。

数据库子系统

建立和维护检索系统中的数据库。
其功能主要包括数据录入、数据库生成、文档更新（增、删、改），及建立索引文档。
对数据库管理，利用的是数据库管理系统（DBMS)
在检索系统中，原始数据经过预处理之后，建立数据库前，先进行标引并创建索引。目前的主流做法是建立倒排索引结构，从而形成倒排文档。检索系统中另一种文档是顺排文档/源文件（类似于顺序文件）
无论是顺排文档，还是倒排文档，管理时都是采用数据库管理方式。

词表管理子系统

功能：管理和维护系统中的词表（也即词库）。
词表就是词汇的集合，可表征数据库中的文献或知识。但对不同的检索系统，词表的形式不同。一种是受控词表，另一种是非受控词表。（词，词间关系，语义关系）
词表的用途：
- **标引用词**。便于查词、选词。（控制词汇的检索系统）
- **检索用词**。以便检索用户准确选词，保证检索效果。

- **词汇查询与输出（服务）**

注：该子系统可独立于其它子系统，也可并入数据库子系统。

提问处理与匹配子系统

功能：处理用户编写并输入的提问，并将其与数据库中存储的数据比较、匹配，其结果输出提供给用户。
例如，检索式：（国防 AND 科技） NOT 贸易.

用户提问>>系统提问（程序/指令）

主要包括四方面的操作：
• 接收提问。通常为检索式或检索词。（事先构造检索策略）
• 提问校验。包括对语法、格式、用词的检查。
• 提问加工。对源提问进行解释性或编译性的加工。如采用波兰展开法等。
•检索匹配。针对接收到的提问，查寻数据库中的信息记录，将满足提问要求的信息记录（检索结果）找出并输出或屏幕显示。
注：通常的信息检索系统都提供检索反馈处理（提问式修改、扩检、缩检）

用户界面子系统

人—机界面(system-userinterface)，提供用户实现检索过程的手段。承担用户与系统间通信和交互功能。
目标：简明、友好（**friendly**）
通常提供：检索命令语言；信息显示方式；信息输入方式。有的还提供智能性接口。
用户界面的发展趋势：自然语言接口(提问、话语输入)、智能接口。

分析上述6个子系统

根据前述信息检索系统的定义，一个计算机化的信息检索系统，通常应具有“信息存储”和“信息查询”两大基本功能。于是，前述6个模块可对应如下：
“信息存储” 部分，包括信息采集、标引处理、数据库管理、词表管理模块；
“信息查询” 部分，包括提问处理（含检索匹配）、用户界面等模块。
两部分之间的桥梁－数据库与索引文档。

信息检索的数学模型

信息检索模型概述

概念

情报检索的本质：情报集合与需求集合的匹配与选择。
信息检索的基本原理是：检索系统对用户信息需求（集合）和系统存储的信息资源（集合）二者间进行的匹配。
select name from /order by/where
create，drop，alter，truncate，comment，rename，insert，update，lock，call

检索模型的作用：精确地描述信息检索的过程和本质；指导信息检索系统的研制与开发。
数学模型：用数学语言描述而得到的一种数学结构。它是对现实世界的某一特定对象作出一些必要的简化与假设，运用适当的数学工具描述而得到。它可解释特定现象的状态和性质，或预测未来的状况，或提供处理对象的最优决策等
现实模型>抽象>数学问题

分析文献检索的本质，它涉及文献集的逻辑表示、用户查询表示、相似性匹配三个主要环节。

信息检索的数学模型: 运用数学的语言和工具，对信息检索系统中的信息及其处理过程加以抽象和描述，表达为某种数学结构。
剖析：对于文献检索系统，从两个方面来抽象处理，一是确定在模型中如何表示两个要素—文献和检索（查询）式；二是确定在模型中如何定义和计算文献和检索式之间的关系。

最简单的检索模型是**单项检索模型**，即由单个的主题词构成检索式。此时，文献集合中的每一篇文献用一个或多个主题词标引。此时，**用户提问**：单个主题词；**系统的响应**：被检出文献/不被检出；**匹配标准**：若提问中的主题词属于某文献标引词集合，那就匹配成功，该文献就作为命中文献反馈给用户；反之就是不匹配，即不命中。
单项检索模型很简单，但效果和适用性较差，往往不能令人满意。特别是当文献集合较大时，查准率往往很低。

二十世纪60年代后期，**布尔检索模型**为一些大型文献检索系统所采用，并逐渐推广应用于各种商业性联机检索系统。
后来，为了弥补布尔检索模型的不足，学者们相继研究提出不同类型的检索模型。

信息检索系统的形式化描述

依据信息检索的内涵，将一个信息检索系统进行形式化描述，表示为一个四元组（**quadruple**）：
S =（D，K，Q，ρ）
D表示检索系统的信息资源集合（如文献集合）
K表示检索系统中索引词的集合
Q表示用户需求集合
ρ表示信息资源与信息需求的匹配计算函数。

D（documents）集合（信息）
K（key words）集合（索引）词表
Q（Query）集合（用户需求）
ρ 函数
剖析：以上元素中，D、K、Q均表示信息，ρ表示处理过程

布尔检索模型（Boolean Model）

布尔模型的数学基础是集合论和布尔代数。
常用的布尔逻辑运算符有三种：逻辑或（or）、逻辑与（and）、逻辑非（not）。

两条基本规则：

（1）系统的索引词集合中的每一个索引词在一篇文献中只有两种状态：出现和不出现。(即索引词的权值仅为二值数据:0 和 1)
（2）检索提问式由三种布尔运算符“and”、“or”、“not”连接索引词而构成。

布尔逻辑运算符

逻辑或（OR）

“逻辑或”，用符号“OR”表示，析取联结词，还可以写成“+”。
A OR B 或者 A+B 的含义：
凡包含词A、或包含词B、或同时包含词A和B的文献，都是命中文献。
[分析]：
对于检索式“A OR B”，假设检索词A命中m篇、检索词B命中n
篇，
“A OR B”结果命中s篇，则：

当A与B不相关时， s=m+n； （无交集）
当A与B有一定相关性时， s<m+n；
当A与B密切相关时， s=Max (m, n)。(一个集合完全包含另一个集合)
综上，Max (m, n) ≤ s≤m+n。
分析表明，“逻辑或”组配可扩大检索范围，增加检索结果数量，提高查全率。

逻辑与（AND）

“逻辑与”用符号“AND”表示，合取联结词，还可以写成“＊”。
A AND B 或者 A＊B 的含义：
只有同时包含词A和B的文献，才是命中文献。 （可用文氏图表示）
[分析]：
对于检索式“A AND B”， 假设检索词A命中m篇、检索词B命中n篇，
“A AND B”结果命中s篇，则：
当A与B不相关时，s=0； （无交集 ）
当A与B有一定相关性时，0<s<m或 0<s<n；
当A与B密切相关时，s=Min (m,n)。 (一个集合完全包含另一个集合)
综上， 0 ≤ s≤ Min (m, n) 。
分析表明，“逻辑与”组配可缩小检索范围，增强检索专指性，保证查准率。

逻辑非（NOT）

“逻辑非”用符号“NOT”或“ANDNOT”表示，否定联结词，还可以写成“－”。
A NOT B 或者 A－B 的含义：
包含第一个检索词A而不包含第二个检索词B的文献，才是命中文献。（可用文氏图表示）
[分析]：
对于检索式“A NOT B”，A命中m篇，B命中n篇，
“A NOT B”命中s篇，则：

当A与B不相关时，s=m；　　当A与B有一定相关性时，0<s<m；
当A与B密切相关时，如果m>n，s=m-n；如果m <n， s=0。

分析表明，“逻辑非”组配可排斥某些检索词的出现，缩小检索范围，保证准确性。

模型描述

布尔检索模型是将一个信息检索系统描述为一个四元组：
S =（D，K，Q，ρ）

D—文献集

D = { d1，**d**2，**d**3，…，**d** n **}**，亦称文献集，**n**—文献篇数(**n** ≧0) 。 D中的元素 d**j (j=1，2，…， n)**为第**j**篇文献，即d**j =（kj1，kj2，kj3，……，kjp），k为索引词。**
前述元素中，D可视为文献库、 dj 可视为文献向量。
例如，假设文献集D中含有两篇文献d1和d2，其中d1含有索引词k1和k2 ，d2含有索引词k1和k3，则**
文献向量 d1=（1，1，0），d2 =（1，0，1）

K—索引词集合

假设一个信息检索系统中存在t个索引词，任一个索引词用ki表示，则全体索引词的集合K可以表示为：
K={k1，**k**2，…，**kt}**，即词库、词表、检索词典。
于是，对于系统中的任一篇文献dj，其主题内容可以用若干个索引词（ki）来表示。

Q—提问集

按照前述检索系统的逻辑结构，用户信息需求可表示为用户提问。从理论上讲，用户的信息需求有潜在真实需求、意识或感知到的需求、表达出的需求、提问（Query）等不同的存在状态。
我们将用户信息需求集合简化为用户的提问集（Q），表示为：
Q={q1，**q**2，…，**q**m**}**，q**i**表示一个用户提问式。
用户提问式是由检索词经布尔逻辑组配构成

ρ—匹配计算函数

依据信息检索原理，信息检索的根本任务就是在信息集合（文献集D）与需求集合（提问集Q）之间基于某种相似性规则进行匹配处理。这里引入匹配计算函数ρ，用来计算任一文献d**i**与任一提问q之间的相似性大小。

笛卡尔积
A✖ B={(x,y)|x∈A∧y∈B}
ρ （q，d）=1，命中

ρ：Q x D → R，R={0，1}。 ρ亦称为映射函数。
Q x D是笛卡尔积，（q，d）是 Q x D的元素；

ρ（q，d）= 1，当文献d满足提问q的要求（匹配成功）
ρ（q，d）= 0， 否则 （不匹配/不命中）
解释：给定提问q ，查询文献集合D，若ρ（q，d）= 1，即文献d满足提问q的要求,则d就是命中文献；反之为不命中。

匹配计算函数 ρ是核心（映射值为0、1两个）。
可借助匹配计算函数来描述检索匹配处理过程。
假设q是检索词的任意布尔组合（由 +、* 、- 组配连接）。按照逻辑学原理，可以将提问q分解为单一检索项或子检索提问式。可表示为两种情形：

q是单一检索项v

定义运算
O(q)=O(v)={d/ρ(v，d)=1}

> O表示运算（Operation），ρ(v，d)=1表示v是文献d的一个元素，即文献d包含v。O(q)表示文献集D的子集合。

q是子检索提问式

q=q1+q2，或 q1＊q2，或q1-q2，而q1、q2可以是单一检索项，或者可以是其任意布尔逻辑组配，则定义运算
O(q)=O(q1+q2)={d/ρ(q1+q2，d)=1}
={d/ρ(q1，d)=1}∪{d/ρ(q2，d)=1}
= O(q1) ∪ O(q2) （集合的“并”）
同理， O(q)= O(q1*q2)= O(q1) ∩ O(q2) （集合的“交”）
O(q)= O(q1-q2)=O(q1) − O(q2) （集合的“差”）
其中，∪、∩、− 分别表示集合的并、交、差运算。

对任一提问q（q是Q的元素），检索响应R(q)= O(q)是文献集合D的一个子集合。我们称之为检索系统对提问q的检索响应。
综上，任一用户提问式均可归结为上述①、②两种情形，利用匹配计算函数，经运算可得出系统对提问的检索响应（即结果文献集，或称为命中文献集，它们是文献集合D的一个子集合）。

例． qi = (k1 AND k2) OR (k3 AND (NOT k4))
该逻辑提问式的含意：要求命中文献d同时含有词k**1**和k**2**，或含有词k**3**但不含词k**4**。
我们可以简化书写为： q**i = (k1 * k2) + (k3- k4))**
按照上述情形归类和检索匹配处理过程，检索系统对提问q的检索响应是：
R(q**i)=** (O(k**1**)** ∩ **O(k2)) ∪ (O(k3) − O(k4))**

布尔检索模型分析

优点

①布尔模型较简洁，可通过简单的逻辑关系来反映检索项之间的联系。
②逻辑运算符较少，便于用户学习和理解。采用的是关键词查询。
③语义表达能力强，能够处理较复杂的检索问题，通过逻辑运算符可将概念间的逻辑关系表现出来，变成计算机运算，实现自动匹配。
④检索提问式比较灵活，方便修改，可调整检索范围。

正因为以上突出优点，使该模型得到了广泛的应用。一些主流的图书馆检索系统和互联网搜索引擎都应用布尔检索模型来构建检索系统。但是，布尔检索模型也存在着一些明显的缺陷。

缺点

①准确匹配标准致使检索的灵活性差。

> 采用二值判断逻辑，认为一篇文献对于某一个提问要么是“相关的”，要么是“不相关的”，这种判断标准严重影响着检索系统的性能。
> 例如： k1 * k2 * k3 * k4，如果文献不含k2，则匹配结果是文献零输出。

②检索中的各组配元的权重未加区分。检索过程中无法反映某个检索项的重要性，也难以反映检索者的主观愿望。

> 认为每一个检索词kj对文献内容贡献程度同等重要，即权重相等。

③检索结果不能按用户希望的重要性排序输出，给用户带来不便。
④构造一个好的检索式并不容易。

布尔检索对用户的素质和语义提取能力要求较高，尤其对复杂的检索问题（课题），提问式不易构造，交互较困难。

向量空间模型

（Vector Space Model）

基本思路

鉴于布尔检索模型“准确匹配”策略造成的检索缺陷，20世纪60年代末期，美国著名学者萨尔顿基于“部分匹配 （**partial matching**）”策略的信息检索思想，在其开发的试验性系统SMART（S**ystem for** M**echanical** A**nalysis and** R**etrieval of** T**exts**）中，提出并采用线性代数的理论和方法，构建出一种新型的检索模型，这就是向量空间模型（**Vector Space Model**，简称**VSM**）。
向量空间模型(**VSM**)**建立在代数理论的基础上，属于一类代数检索模型。
向量空间模型把系统中的文档和检索提问式都看成一组数值向量。

一个向量空间是由一组线性无关的基本向量构成，每一个向量的分量是一个数值。向量具有方向和长度。
模型思想：检索系统中的每一篇文献和每一个提问均用一个等长向量来表示。向量的分量个数

即向量的维数。
例如： 文档**D**中的第**i**篇文献，表示为 **Di=(Ki1,Ki2, ⋯ Kim),**
系统中的第**j**个提问，表示为 **Qj=(Kj1,Kj2, ⋯ Kjm)**
其中m是向量的维数， **K**是索引词。
检索系统中文献与提问的匹配处理**过程，就转化为向量空间中文献向量与提问向量的相似度计算问题。

向量空间模型的基本原理

文献向量与提问向量的含义

对于系统中文献集**D** ，第**i**篇文献 **Di** 可表示为**m-**维向量的形式： **Di=(Ki1，Ki2， ⋯ ，Kim)** 。（即文献向量）
用户的信息需求也被加工、转换为提问向量，表示形式与文献向量类似：**Qj=(Kj1，Kj2，⋯， Kjm)， Qj**表示第**j**个提问。

其中，向量分量**Kin** 表示 **Di** 中第 **n**个索引词，其值代表文献 **Di** 能力的大小，及权重（体现该词在文献**Di** 中的重要程度）；向量分量**Kjn**表示**Qj**中的第**n**个检索词。 **Kin** 和 **Kjn** 的取值范围是在0和1之间。

系统中的每一篇文献就表示为**m-维文献向量，每一个提问也表示为**m-维提问向量。**m**是系统中索引词的总个数。
例如，假定m的值是6
文献向量 Di = (0.5，0.0，1.0，0.7，0.9，0.0) ，提问向量 Qj = (0.6，0.0，0.0，0.8，0.9，0.0)。 其中，向量的分量值反映词的权重。

匹配函数及相似度阈值的确定

文献与提问之间的相似度（**Similarity**），即二者间的相关程度大小，可以由它们各自对应向量在**m**-**维空间中的相对位置来决定。
匹配函数一般采用余弦函数法。（向量夹角的余弦值）
余弦函数法的实质：计算m-维空间中某一个文献向量与某一个提问向量之间夹角的余弦。
夹角越小，说明两个向量越靠近，余弦函数值越大，对应文献与提问条件越相关，命中的可能性就越大。反之就越不相关。

由于系统不可能将所有的相关文献全部呈现给用户，而只能将相似度较高的文献作为检索结果，这就需要设定一个阈值（**threshold**），用 表示。凡相似度值大于k的文献，都将作为检索结果提供给用户。

向量空间模型的检索匹配体现了一种“部分匹配”的策略。

在实际检索中，当全部文献向量与某个提问向量的相似度都计算完毕后，系统就把相似度值超过某一规定阈值的文献按相似度大小排序输出。

（1）该模型体现了一种柔性的信息检索。多数情况检索结果介于命中和未命中之间，而不是单纯地确定一篇文献要么被命中、要么不命中。

（2）相似度阈值设定后，对大于阈值的，则保留该文献查询结果；对小于阈值的，则过滤（舍弃）掉。

向量空间模型的分析

优点

（1）采用部分匹配策略，提高了检索的弹性。
（2）标引词和文献的相关程度可在0和1之间取值，改变了布尔模型只有0和1两种结果的简单判断。使得标引者和检索者都可以比较灵活地定义标引词和文献的关系深度，克服了布尔检索模型僵化的缺点。
（3）有了相似度的标准，就可以从量的角度来判断文献命中与否，使检索更趋合理。检索结果可按照与提问的相似度高低排序输出，便于用户对检索结果数量进行控制与调整。

缺点

（1） 对向量的相似度计算工作量大，算法复杂度较高。因为对任何一个提问，都需要计算全部文献集中的每一篇文献。
（2）文献向量中各分量的值（索引词权值）较难确定。确定权值具有随意性（主观性），将影响检索质量。

向量空间模型为揭示情报检索的基本原理作出了重要贡献。它具有有效的匹配算法设计，可取得较为满意的处理结果，使得基于该模型的研究思路大为流行，并大量应用于文本处理领域。

向量模型中采用的相似度计算，除了可以计算文献和提问的相似度，还可计算文献之间的相似度，即计算**Di** 和**Dj**的相似度**Sim（Di，Dj**),**从而使属性相似的文献尽量聚集在一起，从而提高检索效率。

文本信息处理应用领域

文本聚类（**Text Clustering**）；
文本检索（**TextRetrieval**）；
文本分类（**Text Categorization**）；
文本过滤（**Text Filtering**）；
文本浏览与可视化（**Text Browsing and Visualization**）等。

其它检索模型简介

概率检索模型*

概率检索模型是基于概率论（**Probabilistic theory**）原理，以解决不确定性的信息检索问题。该模型认为：给定文献和给定提问之间存在某种相关概率。

指导思想：给定一个用户提问，则检索系统的数据库中存在着一个与该提问相关的理想命中结果集合。

模型的基本原理：利用概率论思想，通过赋予标引词某种概率值，来表示这些词在相关文献集合或无关文献集合中的出现概率，然后计算某一给定文献与某一给定提问相关的概率，最后系统据此作出检索判断或决策。（基于**Bayes**决策理论）

模型的特点：理论上严密，检索到的文献可按相关概率值排序输出。但计算开销大，且相关参数估计难度大。

模糊集合模型

从信息检索的本质上讲，文献与提问之间的匹配处理实际上是一个近似的过程。在检索系统中，词与文献、提问与文献之间的关系都可视为是“模糊”的，即具有不确定性、或在一定程度上相关

模糊集合模型建立在模糊集合论的基础上。模糊集合论的中心思想是把隶属函数和集合中的元素结合在一起。认为不同元素对于同一集合具有不同的隶属程度，从而引入隶属度概念，在**0**～**1**之间取值，即**[0，1]**。

为表示文献与标引词相关的程度，采用隶属度。如d1={(t1，0.5)，(t2，0.8)}；d2={(t1，0.9)，(t2，0.1)}，t1和t2表示标引词，后面的数字表示文献d1和d2对它们的隶属度。

用户通过一个布尔型的提问表达式来阐述他的需求，并指定所需文献对检索词的隶属度（权值）。计算文献与提问相关的过程类似于布尔模型运算匹配过程。检索结果可按文献的权值大小排序输出。

位置检索模型

位置检索：针对自然语言文本中检索词与检索词之间特定位置关系而进行的检索。例如，两个检索词相连或相邻、或在同一句中、同一字段中的检索，从而使检索出的文献更符合用户要求，提高查准率。可用于文本（段）检索。

为实现检索词在检索结果中应满足的位置，需要使用一些位置运算符。如：
(W)算符：在检索提问式中，两个检索词必须在文本信息中按照先后顺序紧挨着出现，两个检索词之间可以有一个空格、一个标点符号或者一个连接号，而不得夹有任何其他单词、字母或汉字。(nW)算符类似。 （W，是With的首字母）

(N)算符：在检索提问式中，它所连接的两个检索词必须紧密相连着出现，两个检索词之间除了可以有一个空格、一个标点符号或者一个连接号外，不得夹杂其他任何单词。但是，与(W)的区别是
(N)算符两侧的检索词出现次序可以颠倒。(nN)算符类似。（N，是Near的首字母）

(S)算符：要求参加检索运算的两个词必须在同一自然句中出现，其先后顺序不受限刺（同句检索）。**S**为“**sentence**”一词的首字母。

(F)算符：表示在此算符两侧的检索词必须同时出现在数据库记录的同一个字段中，词序可变（**同字段检索**）。F是单词field的首字母。

限制检索模型

限制检索：一些缩小检索范围或约束检索结果的检索。
限制检索的形式有很多，最主要的限制检索是通过限制检索词在命中结果记录中的出现位置（主要指记录的不同字段位置）来实现的，故也称之为“字段检索”。
例如，检索文献篇名中含有“**robot**”的相关文献，可用检索式
robot within TI
又如，查找 作者**wang wei** 写的文章，可用检索式
au = wang wei
也可通过语种、出版国家、出版年代等的字段标识符，来限定检索范围。