# 信息资源管理导论重点

## 名词解释

### 1.普莱斯定律

在洛特卡定律的基础上，普赖斯进一步研究了科学家人数与科学文献数量，以及不同能力层次的科学家之间的定量关系，提出了著名的普赖斯定律和一些其他重要结论。在《小科学，大科学》一书中，普赖斯写道：“科学家的总人数，大致是按杰出科学家人数的平方增长的。”所谓**普赖斯定律（Price Law）**，即**科学家总人数开平方，所得到的人数撰写了全部科学论文的 50%**。

如果设最高产的那位科学家所发表的论文数为 nmax，将科学家们发表论文的总数记为 x(1， $n_{max}$)，则普赖斯定律可用下式表示：

$$(1/2)x(1, n_{max}) = x(m, n_{max}) = x(1, m)$$

### 2.文献半衰期

文献的半衰期，是指某学科领域现时**尚在利用的全部文献中的一半是在多长一段时间内发表的**。

### 3.信息组织

信息组织即**信息序化或信息整序**，是指利用一定的科学规则和方法，通过对信息外部特征、内容特征或其它特征的描述与序化，实现对信息资源进行选择和整理，从而到达充分利用的目的。**原则**：客观性、系统性、实用性、

### 4.元数据

元数据被定义为：**描述数据的数据，对数据及信息资源的描述性信息**。

元数据（Metadata）是描述其它数据的数据，或者说是用于提供某种资源的有关信息的结构数据（structured data）。元数据是描述信息资源或数据等对象的数据，其**使用目的**在于：识别资源；评价资源；追踪资源在使用过程中的变化；实现简单高效地管理大量网络化数据；实现信息资源的有效发现、查找、一体化组织和对使用资源的有效管理。（百度百科）

### 5.标题法（标题词法）

**用经过规范化处理的自然语言语词及语组来逐一表达主题概念**

标题词法特点

- 因为标题之间的顺序关系是预先组配好的，先组配式的标题在标引和检索时直接使用，不易混乱；
- 因为标题法以事物为中心来集中与该事物有关的文献，适合于从主题出发进行检索，易于查全一项事件的文献；
- 直观、易掌握，检索速度快，对新事物、范围细小的问题容易反映出来，补充修改也比较容易。

## 6.单元词法（元词法，Uniterm ）

用来标引信息资源主题的、最基本的、最小的词语单元

单元词法特点

- 最基本的特点是**概念组配**。
- **优点**：强调词汇的单元化和后期组配，因此提高了主题法的灵活性；
- **缺点**：由于它过分强调词汇单元化，词汇处理方法又不甚合理，易发生错误组配，误检率较高，故实用性不是很好。

## 7.叙词法（主题词法）

**叙词**，也称主题词，是**经过规范化处理的、以基本概念为基础的表达文献主题的词或词组**。**叙词法**，就是通过精选的自然语言语词的概念组配来表达主题的方法。

**叙词法特点：**

- **直观性**：直接以规范化了的自然语言叙词作为标识符号。
- **专指性**：直接从文献论述和研究的具体对象和问题出发进行选词，并采用叙词组配来描述主题，所以不论文献主题如何专深，都可根据需要直接选作叙词或通过组配加以表达。
- **适应性强**：对不断出现的新事物、新学科、新概念和新的研究课题，叙词法能随时加以增删和修改。
- **迅速准确**：对叙词主要采取字顺排列方式，因此查找方便、迅速准确。

## 8.关键词法（Keyword）

**直接以文献中的语词来表达主题概念**

关键词法特点：

优点：关键词法不受词表控制，快捷简便，适于用计算机组织和检索文献信息；

缺点：由于关键词法的词语不规范，影响了信息资源的查全率和查准率。

## 9.词干化

词干化是**去除词尾变化或是有时将派生词变回它们的词干——基本形的过程。**比如，一种词干化算法可能会将 Riding 和 Rides 转化为 Ride。词干化有助于提高结果召回率，但是会对准确率造成负影响。

词干化处理就是**把一些名词的复数去掉，动词的不同时态去掉等等类似的处理**。对于切词得到的英文单词要进行词干化处理，主要包括将名词的复数变为单数和将动词的其他形态变为基本形态。

时态问题统归到词的原型，单复数统归到词的原型，**处理单复数和时态的问题就是词干化的过程**；合并近义词，在西文中词源归一化的问题

## 10.自组织

网络中的信息由于用户与用户之间、用户与网络其他要素之间的交互性、相关性、协同性而形成的特性结构和功能，使得开放信息网络具有了非**平衡系统的自组织现象**。

信息自组织驱动力

**宏观层面**

从宏观层面而言信息自组织的外驱动力是信息系统存在的外部竞争，内驱动力是信息系统本身进化的要求。

**微观层面**

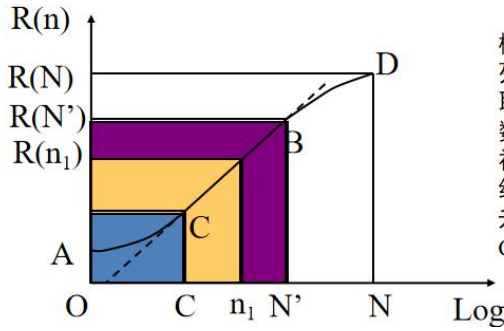
从微观层面而言，信息自组织的外驱动力包括技术性驱动力和社会性驱动力，即信息技术和信息用户的创造性和参与性意识，内驱动力是信息系统各子系统之间及其内部元素之间的相互作用。

## 11.离散分布规律——布拉德福定律

- 亦称“文献分散规律”
- 把期刊分为专门面对这个学科的核心区、相关区和非相关区
- 分类标准为刊载某学科专业论文的数量
- 各个区的文章数量相等
- 核心区、相关区，非相关区期刊数量成 **1:n:n²** 的关系

布拉德福定律图形描述

- 假设一定时间内（通常为一年）共有N种期刊刊载了某学科的论文（简称为“相关论文”）K篇，将这N种期刊按照所载“相关论文”的数量降序排列，然后，以期刊累积数量的对数（lgn）为横坐标，以相应的“相关论文”累积数量（R(n)）为纵坐标作图如下：




横坐标取期刊按载文量递减排列时的顺序号n的对数，纵坐标取1至n号期刊所载论文的累积数，绘制出的曲线称为布拉德福分散曲线。 布拉德福分散曲线由三部分构成，先是一段上升的曲线AC，然后是一段直线CB，最后是下垂的曲线。

19

- 维克利对布氏定律的推论
  - 针对分区进行修正
  - 分区不同，比例系数就要发生相应的变化
- 修正后的布氏定律

$$n_1:n_{1-2}:n_{1-3}:\cdots : n_{1-m}=1:V:V^2:\cdots : V^{m-1}$$

公式中：

$n_{1-k}$（k=2,3,… ,m)——第一区到第k区的期刊累计数量

m——划分的区域数

V——分散系数（或称为维氏系数）

维克利的论证和补充，使布拉德福文献分布的图像与定律在结构上得到了统一，丰富了布氏分布理论的内容，使其在形式上趋于完整，为布拉德福定律的确立和发展做出了重要贡献。

## 布拉德福定律的理论解释

**文献为什么“离散”?**

**科学统一性原则**:每一个科学学科都或多或少，或远或近地与其他任何一个学科相关联。因此，属于某学科的文献，不仅仅会出现在这个学科的专业期刊上，而且也时时可能出现在其他学科的期刊上。

**文献为什么“集中”?**

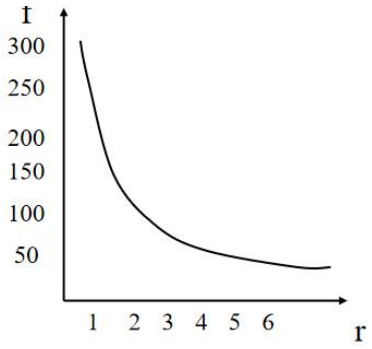
**成功产生成功的原则( the success-breeds-success principle）** :一种期刊的文献量越大、质量越高的期刊，作者就越愿意将自己的文章刊登在这些期刊上，形成了一种“堆加效应”。

**“文献分布的集中与离散规律”**

## 齐普夫定律

如果将一篇较长文章(约 5000 字以上)中每个词按其出现频次递减排列起来(高频词在前，低频词在后)，并用自然数给这些词编上等级序号，出现频次最高的为 1 级，其次为 2 级……这样一直到 D 级，如果用 f 表示词在文章中出现的频次，用 r 表示词的等级序号，则有

Fr=c(c 为常数）




- 齐普夫分布曲线
  - 以等级序号为横坐标
  - 以相应的频次为纵坐标

齐普夫的表达仅适宜于中频词的情况，高频与低频词与该表述偏差较大

### 齐普夫定律的修正——朱斯的双参数公式

- 美国语言学家朱斯对齐普夫的单参数词频分布律提出了修正
- 齐普夫定律中r的负指数应该是一个参数，而不是一个常数
- 双参数公式如下：

$$P_r=C*r^{-b}\ （b>0，\ C>0）$$

### 齐普夫定律的修正——芒代尔布罗的三参数公式

- 运用信息论原理和概率论方法来研究词的频率分布定律
- 三参数频率分布定律:

$$P_r=C*(r+a)^{-b}\ \ (0<=a<1，\ b>0，\ C>0)$$

## 12.六度分离

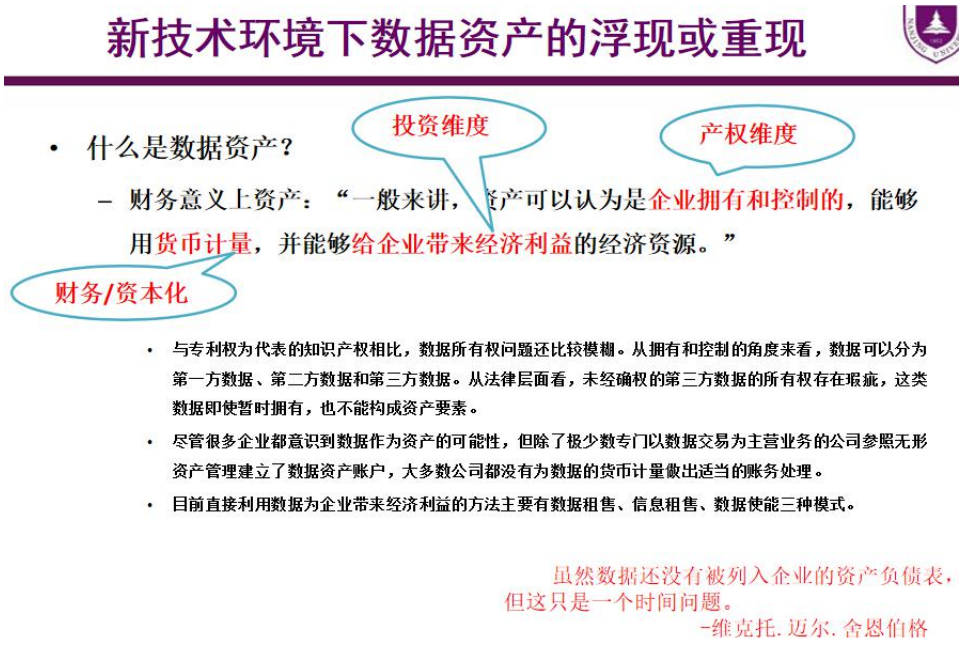
**小世界现象**是指人与人之间的联系可以通过有限个个体连通，也称为**“六度分离”理论**，即你和任何一个陌生人之间所间隔的人不会超过 6 个。

你和任何一个陌生人之间所间隔的人不会超过五个，也就是说，最多通过五个人你就能够认识任何一个陌生人。生活在这个世界上的每个人平均只需要通过 6 个中间人就能与全世界每一个人建立联系。

## 13.数据资产和大数据

**财务意义上资产：**“一般来讲，资产可以认为是企业拥有和控制的，能够用货币计量，并能够给企业带来经济利益的经济资源。”

**数据管理维度：**并不是所有的数据都是资产，**只有可控制、可量化、可变现的数据才能成为资产**。数据资产的**特点**包括：虚拟性、共享性、时效性、安全性、交换性和规模性。其中，**共享性**尤为重要。




## 大数据

## 简答

### 1.大众分类法 Folksonomy=Folks+Taxonomy

Folksonomy 是一个合成词，是指大众自发参与的分类方法，也是由社会化标签服务中最具特色的自定义标签功能衍生而来。Folksonomy 的大众化、自由化和社会化的特征使其适用于网络信息资源的组织和管理，尤其是用户生产的各类信息资源的组织和管理。

Folksonomy 分类方法一般由网络信息用户自发为某类信息定义一组标签进行描述，并最终根据标签被使用的频次选用较高频次的标签，作为该类信息类名，的一种为网络信息分类的方法。

与传统方法相比， Folksonomy 是借助于用户的群体智慧，而非信息组织加工者的专业知识，来揭示信息资源的特征。

特点：

**自由灵活**：以自定义的自由此为数据资源对象进行标注与分类

**共建共享**：用户对内容标注后，其他人可浏览查阅或自行添加新标签

**动态更新**：随时反映用户的关注“热点”与“走势”

大众分类法的缺点

**缺乏层次性（lack of hierarchy）**

- 一种平面的分类方式，很难使用它来揭示复杂的关系

**表达概念的模糊性**

- 缺乏语义精确性
- 缺乏同义词控制
- 词的多义性问题
- 用户标签五花八门，加重系统负担，降低分类准确性

### 2. 网络活动中用户使用信息组织的问题


武汉大学资深教授马费成发言

## 微博客信息组织方式

**大众标注**

微博客用户可以通过大众标注方式揭示并与他人共享信息。

**内容聚合**

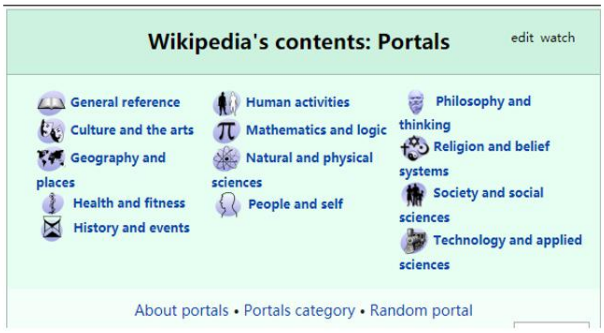
微博客用户通过内容聚合方式动态聚集相关主题的信息。内容聚合是博客站点之间共享信息内容的一种简易方式，是基于XML规范的应用。

**引用通告**

微博客用户可以通过引用通告方式推送并聚集关于同一主题的讨论信息。

**内容挖掘**

微博客用户可以通过digg（挖掘，即链接、评论）评论和推介信息。




**英文维基百科的分类框架:上部有概览、主题、知识纲要、导航、特色内容、术词表、分类和字顺索引**

文化中的各国文化子类:点击子类名,显示全部内容,点击“➡”，显示细目。
与分类索引中的类目不一样。灵活的体系。

特色内容导览




多重分类方式:分别按字顺、音序、学科、时间、相关等多维方式分类。

网络中的信息由于用户与用户之间、用户与网络其他要素之间的交互性、相关性、协同性而形成的特性结构和功能，使得开放信息网络具有了非平衡系统的自组织现象。

一个用户广泛参与的知识系统，知识增长过程实质上是输入信息使知识结构由无序走向有序，或从一种有序结构演变为另一种有序结构的过程。

## 3. 信息交流与要素

### 信息交流(Information Communication):

社会活动中的认知主体借助某种符号系统，利用某种传递通道，在不同时间（历时）和空间（共时）上实现的信息传输和交换行为。

认知主体:人或由人组成的机构、组织;包括“发送”和“接受”双方

实现条件:符号系统、传递通道

基本前提:不同的时间、空间

### 信息交流要素

**①信息发送者（信息传递者、信息生产者、信源、Sender)**:信息的初始来源，是信息传递链上的初始环节;

**②信息接收者（信息接受者、信息利用者、信宿、Reciever) :**信息的最终接受者或利用者，是信息传递链上的最终环节;

**③交流通道（信息渠道）**:传送和交换信息的媒介和工具;

“媒介”:信息生产、接收和传递中所依托的物质载体，如印刷文献、电子书、缩微胶片、信息网络等等;

“工具”:信息生产、接收和传递中所使用的物化工具，如计算机、电视机、录音机、摄像机、缩微阅读器等等;

**④符号体系:**传递信息的符号元素及其相互之间的联系与组织方式、规则;

**“符号”:**一种代表思想的通用记号或标志，是用以表达思想、进行指挥或者表示愿望的一种标记、行为或姿态，如语言、文字、手势、表情、信物、烽火狼烟、旗语、计算机语言等等;

**⑤知识信息库:**人脑知识信息的总称，是信息交流的最根本来源和最终极的归宿;

**⑥支持条件:**保障信息交流得以实现的自然、技术和社会条件。

- 自然条件，如光、声、电、空气等;
- 技术条件，如各种通讯技术、存贮、处理技术等;
- 社会条件，如法律、政策、经济条件、信息机构及相关组织的建立等。

## 4. 共时信息交流和历时信息交流

### 含义

共时信息交流（横向信息交流）:发送者和接收者在同一时间层面;

历时信息交流（纵向信息交流）:发送者和接收者在不同时间层面;

### 主要功能:

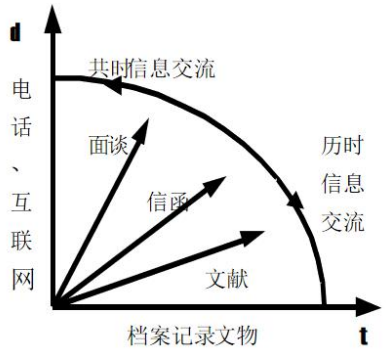
共时信息交流:消除交流的空间障碍;

历时信息交流:消除交流的时间障碍;

### 交流通道

| | 共时信息交流 | 历时信息交流 |
|---|---|---|
| 交流手段 | 互联网、传真、电话、电报、广播、电视、邮政、身势、旗语、钟、鼓、灯、烽火、口语、实物等 | 刻制光盘、拷贝磁盘、录音、录像、照相、绘画、文献、档案、古迹、文物、口语等 |

相互关系




## 5. 信息组织原则和方法

信息组织的原则：客观性、系统性、实用性。

**信息的外部特征**

- 信息的外部特征是指信息的物理载体直接反映的信息对象，构成信息的外在的、形式的特征
- 信息载体的物理形态
- 题名、作者、出版或发表日期
- 流通或传播的标记等方面的特征；

信息的内容特征　信息的内容特征就是**信息包含的内容**，它可以由**关键词、主题词或者其他知识单元**表达。

信息标引语言：

**按照信息组织的思想方法**，标引语言主要可以分为分类语言与主题语言。

分类语言，是用分类号来表达主题概念，依据知识分类方法将主题概念逐层细分为类目体系，类目体系是分类语言的基本形式，而系统性是其主要特征；

而主题语言则主要是通过受控语言来表达主题概念，并按字顺或一定规则排列主题概念的方法，参照显示是其基本形式，以事物为中心的直接性是其主要特征。

**按照标识语言的规范和受控程度**，标引语言可以分为受控语言与自然语言。

受控语言是指信息组织的标识语言以及索引词汇在使用前经过优选和规范化处理，具有统一的规范体系，因此也称为规范化语言。比如单元词、叙词和体系分类语言等；

而自然语言是指标引词汇和索引词汇直接来源于文献或信息源本身，比如文献中的关键词。

信息标引的类型

**分类标引**：是指对文献进行主题分析，用特定的分类语言(分类法)表达分析出主题，赋予文献分类检索标识(分类号)的过程。

**主题标引**：是指对文献进行主题分析，用主题语言（主题法）表达分析出主题，赋予文献主题检索标识(标题词、叙词等)的过程。

总结：不同视角下的分类法和主题法

| 视角 / 方法 | 检索语言 | 检索标识 | 标引工具 | 标引方法 |
|---|---|---|---|---|
| 分类法 | 分类语言 | 分类号 | 分类表 | 分类标引 |
| 主题法 | 主题语言 | 主题词 | 主题词表 | 主题标引 |

## 信息组织的基本方法

### 一、分类法

**分类**：就是按照事物的性质、特点、用途等进行区分和聚类，并将聚类结果按照一定次序予以组织的活动，是一种系统认知事物的方法。

**分类法：**一种依照内容特征将信息资源分门别类、系统组织和揭示的方法。一般采用一定的标记符号作为排序工具。

**类型**

**（1）等级列举式分类法** ：将所有的类目组成一个等级系统。如，《杜威十进制法 DDC》、《美国国会图书馆分类法》。

**（2）分面组配式分类法**：在类目之间完全采用分面结构，将交献的内容分析为若干个因素、从分面寻找相应的类号，并按照一定的次序将其排列组配成一个完整的分类号。如，《冒号分类法》。

半分面分类法：以等级列举式的类目基础，在类目拓展方面采用分面组配的方法，实现等级列举式类表与分面等标引的功能。

### 二、主题法

**主题法**是直接以表达主题内容的语词作为检索标识，以字顺为主要检索途径，并通过参照系统等方法揭示词间关系的标引和检索信息资源的方法。

主题法的特点：

- 以特定的事物、问题、现象，即以主题为中心集中信息资源。
- 直接以词语为信息资源检索标识。
- 以字顺为主要检索途径。
- 往往通过详尽的参照系统等方式揭示主题词之间的关系。

（1）标题法
（2）单元词法
（3）叙词法
（4）关键词法

### 三、分类主题一体化

分类法系统性强，适用于族性特征的标引和组织；
主题法直观明确，适用于从特性特征的标引和组织。

而在人们的信息资源利用过程中，既有族性检索的需求，也有特性揭示的目标。因此，产生了两种方法互为渗透的**分类主题一体化词表。**比如巴希特 1969 年编制的《教育检索词表》和**《中国分类主题词表》**等。

### 四、分类法和主题法的优缺点

分类法

**优点**：分类标记的成分一般为数字或字母，通用性好，不受语言、国别的影响，是一种世界性的通用信息组织方法。

**缺点**：直观性较差，需要有专业的知识背景，普通用户的使用具有一定的局限性。

主题法：直观明确，适用于从特性特征的标引和组织

## 6. 信息的概念；认识论、本体论、三要素

**什么是信息？**

**申农**：信息是用来减少随机不确定性的东西，即负熵

**信息资源管理学界认为：**

信息是数据处理的最终产品，即信息是经过采集、记录、处理，以可检索的形式存储的事实或数据。

**信息的基本功能**

1、信息是人类生存的前提

2、信息是人类发展必需的重要资源

没有物质，系统就无形体，没有能量，系统就没有活力，没有信息，系统就没有灵魂。

3、信息是人类一切智慧和知识的源泉

**认识论层面的三要素**

**研究信息传递主要是涉及语法信息;研究信息的意义主要涉及语义信息;而研究信息的效用则主要涉及语用信息。**

**语法**是指符号与符号之间的关系，是书面语言、口头语言、计算机编码语言或行为语言的规则。语法信息也被称作技术信息，它是研究符号与符号的形式关系，研究它们的编码、传递、重复和再现等，并不考虑这些符号的实际意义和效用。语法信息具有客观的本性，它是信息最基本的层次;人们总是通过实践感知事物的表征，获取原始材料，产生语法信息。

**语义**是指符号与实体之间的关系，是符号所表达的内容和含义。语义信息与语法信息相反，它表明信息的内容含义以及在逻辑上的真实性、准确性和可靠性。人们对接触到的事实材料(语法信息)加以理解，把它与它所代表的实体联系起来，进行分析比较形成某种概念，这就是语义信息。它是对事物运动状态的陈述或表征，其目的是保证接收者得到信息的实际内容。语义信息有客观的一面，又有主观的一面。对信息含义能否正确理解以及理解的深广程度，与每个人的认识水平、理解能力等主观因素有关。

**语用**是指符号与使用者之间的关系，指符号所表达的内容和含义对使用者的效用。语用信息是指信息对接收者的效用，即信息的实用性和价值。信息没有这一层次也就失去了存在的意义。语用信息比语义信息更加依赖接收者，具有更明显的主观色彩，而且它与时间有密切的关系。（来源于网络）

**Q2:本体论与认识论的信息层次**


**核心：本体论和认识论**

本体论：客观

认识论：主观

信息：本体论到认识论的过程

信息的认识层次

例子：


| 层次 | 信息内容 | 描述的问题 | 认识论层次 |
|---|---|---|---|
| 1 | 迹象 | 什么? | 语法信息 |
| 2 | 事实 | 是什么? | 语法信息 |
| 3 | 知识 | 为什么? | 语义信息 |
| 4 | 智慧 | 怎么办? | 语用信息 |

## 7. DIKW 模型

数据，转化为信息，升级为知识，升华为智慧，这个过程是**信息的管理过程**，让信息从庞大无序到分类有序。

**（1）数据（Data）：**

- **数字、文字、图像、符号等的集合，通过原始的观察和度量得到。**
- 数据是最原始的素材，未被加工解释，没有回答特定的问题。
- 数据是载荷或记录信息的按照一定规则排列组合的物理符号。它可以是数字、文字、图像，也可以是声音或计算机代码。
-

**（2）信息（ Information ）：**

通过某种方式组织和处理数据，分析数据间的关系，数据就有了意义。这种**经加工处理后有逻辑的数据，称之为信息。**

信息是已被处理，具有逻辑关系的数据，是对数据的解释，这种信息对接受者具有意义。

回答的问题：Who, What, Where, When

人们对信息的获取只能通过对**数据背景和规则**的解读。背景是接收者针对特定数据的信息准备。信息是数据载荷的内容，对于同一信息，其数据表现形式可以多种多样。

**数据+背景=信息**

**（3）知识（Knowledge）：**

知识是从**相关信息中过滤、加工、提炼而得到的有用资料。(有用的信息**）

特殊背景或语境下，知识将数据与信息、信息与其在行动中的应用之间建立

有意义的联系，它体现了信息的本质、原则和经验。

回答的问题：How

知识是信息接收者通过对信息的提炼和推理而获得的正确结论，是人通过信息对自然界、人类社会以及思维方式与运动规律的认识与掌握，是人的大脑通过思维重新组合的、系统化的信息集合。知识经过推理和分析，可能产生新的知识。

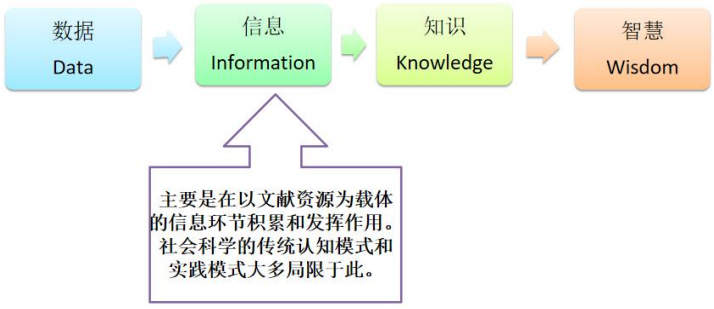
**信息+经验=知识**

**（4）智慧（Wisdom）：**

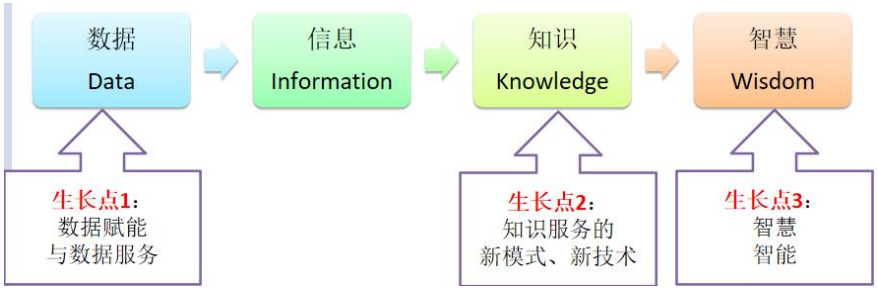
智慧是**人类所表现出来的一种独有的能力**。主要表现为：收集、加工、应用、传播知识的能力，以及对事物发展的前瞻性看法。

在知识的基础之上，通过经验、阅历、见识的积累，而形成对事物的深刻认识、远见，体现为一种卓越的判断力。

回答的问题：Why




- **事实**通过描述成为数据
- **数据**通过特定的背景成为信息
- **信息**通过加工吸收提炼成知识
- **知识**用于特定目的被激活成为情报，解决问题成为**智能**




贯穿数据、信息、知识、智慧链条上的生长点，实现转换尤为重要。

## 8. 数据治理的问题

**状态：**评价对象和标准，评价的是一种对信息有序利用，最大化信息利用的状态；
**手段：**多元介入多元处理的方式
治理相对于管理最大的特点是管理要素和工具的采用是多元的
目标是使数据达到有序并且高效利用的**状态；**
**导向**
**对内：**要求数据达到序化、资产化（诉求非常高）、可重用化；
资产化既包括跟内部业务结合产生的**效能和效用**
对外：产生价值

实现**效能和效用**指标

不同类型企业的侧重点不一样，商业性企业效能性指标一些；

但**政府**必须两个指标都要达到：电子职务很好的去运转；政府的数据能支撑公众很好地服务。

要达到这样的目标，不是一次能完成的，是一项复杂的系统流程，有许多法律、政策、技术、管理因素等介入、

数据治理的最终目标是提升数据的价值，数据治理非常必要，是企业实现数字战略的基础，它是一个管理体系，包括组织、制度、流程、工具。

## 数据治理

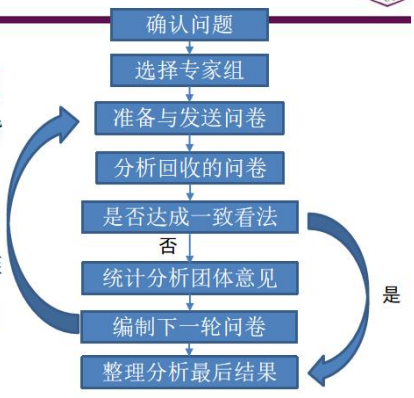
- 数据治理的**驱动力**最早源自两个方面：一是为了满足外部监管和合规的需要；二是内部风险管理的需要，包括财务做假、敏感数据涉密等。
- IBM对数据治理的认识和定义为：数据治理是根据企业的数据管控政策，利用组织人员、流程和技术的相互协作，使企业能将数据作为企业的核心资产来管理和应用的一门学科。**好的数据治理框架应该包括两部分内容：**一是数据治理的**保障机制**，二是数据治理的**核心领域**。
- 数据治理的**核心**是为业务提供持续的、可度量的价值。
- 数据治理的**根本目的**是发挥数据的价值，是一项系统工程， 不仅涉及技 术，还涉及政策、环境、法律 法规、公共管理等多个方面。
- 国外数据治理主要关注3个方面：
  - 第一是重视和完善数据基础设施的建设；
  - 第二是推动和加大政府数据开放；
  - 第三是吸引和鼓励企业和公众对政府大数据进行开发利用。
- 主要议题有：
  - 政府部门内部的数据共享与融合：元数据管理策略、数据集成
  - 政府数据开放，
  - 政府数据的市场化利用，从而形成政府大数据的产业链和价值链

## 9. 德尔菲法

### 德尔菲法

德尔菲法是一种**以集体访谈为基础的预测性调查方法**，但在操作时，德尔菲法还是“背靠背”，以个体方式进行。

该方法始于20世纪40年代，由美国著名的咨询机构兰德公司首先提出，由于其预测的准确性较高，后来人们就以德尔菲（**Delphi**）命名。

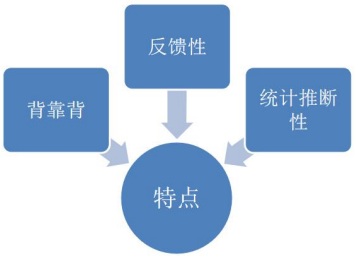



德尔菲法实施流程

**流程：**

- 调查机构将研究问题写成调查提纲，分别送给经过专门选择的专家，请专家用书面的形式对问题做出回答；
- 入选专家独立的给出自己的回答并反馈给研究机构；
- 研究机构将所有专家的意见汇总进行定量分析并将结果反馈给各个专家；
- 专家们根据反馈资料，重新考虑原来的意见，然后将最新的意见反馈发给研究机构；

**特点**




**（一）匿名性**

因为采用这种方法时所有专家组成员不直接见面，只是通过函件交流，这样就可以消除权威的影响。这是该方法的主要特征。匿名是德尔菲法的极其重要的特点，从事预测的专家彼此互不知道其他有哪些人参加预测，他们是在完全匿名的情况下交流思想的。后来改进的德尔菲法允许专家开会进行专题讨论。

**（二）反馈性**

该方法需要经过 3～4 轮的信息反馈，在每次反馈中使调查组和专家组都可以进行深入研究，使得最终结果基本能够反映专家的基本想法和对信息的认识，所以结果较为客观、可信。小组成员的交流是通过回答组织者的问题来实现的，一般要经过若干轮反馈才能完成预测。

**（三）统计性**

最典型的小组预测结果是反映多数人的观点，少数派的观点至多概括地提及一下，但是这并没有表示出小组的不同意见的状况。而统计回答却不是这样，它报告 1 个中位数和 2 个四分点，其中一半落在 2 个四分点之内，一半落在 2 个四分点之外。这样，每种观点都包括在这样的统计中，避免了专家会议法只反映多数人观点的缺点。

## 计算

### 倒排文档（信息检索部分）

- 文档表示（Document Representation）

– 标引词与文档表示（倒排文档）：令 t 表示文档集里所用不同标引词的数目，$K_i$ 表示一个标引词，$K=\{K_1, K_2, K_3, \cdots K_t\}$ 表示所有标引词的集合，对于文档 $D_j$ 中存在的标引词 $K_i$，其权重 $W_{ij}>0$；对于文档 Dj 中没有的标引词 $K_i$，其权重 $W_{ij}=0$。这样就可以将文档 Dj 表示成一个向量 $Dj=(W_{1j}, W_{2j}, W_{3j} \cdots W_{tj})$，向量 $D_j$ 的第 i 维就对应项 Ki 在文档 Dj 中的权重。

| | 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 |

## 需求表达（User Demand Presentation）

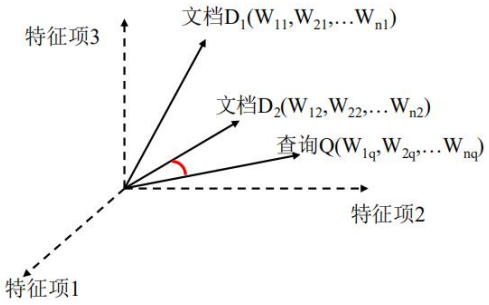
• 用户查询中的标引词一般具有权重特征，设 Wiq 是用户检索提问式 Q 的标引词 Ki 的权重，且 Wiq≥0，则用户查询向量 Q 被定义成：Q=（W1q，W2q，W3q……Wtq）

衡量文档和查询的相关度转化成计算文档向量和查询向量之间的相似度。一般使用文档向量和查询向量之间的夹角余弦值来计算它们之间的相似度。

文档向量空间的表示：

| $W_{ij}$ | $K_1$ | $k_2$ | … | $K_n$ |
|---|---|---|---|---|
| $D_1$ | 0 | 1 | … | 0 |
| $D_2$ | 1 | 0.8 | … | 0.5 |
| … | … | … | … | … |
| $D_n$ | 0.2 | 0 | … | 1 |

文档向量空间模型：



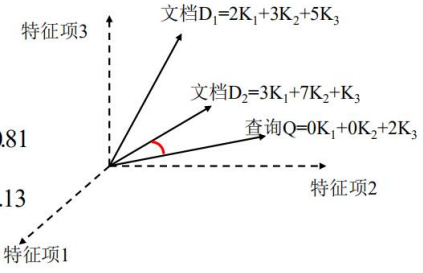

文档和查询之间的相似度Sim可以表示如下：

$$Sim(D_i,Q)=\cos\theta=\frac{\sum_{j=1}^{n}W_{ij}\times W_{iq}}{\sqrt{\left(\left(\sum_{j=1}^{n}W_{ij}^2\right)\left(\sum_{j=1}^{n}W_{iq}^2\right)\right)}}$$

文档和文档之间的相似度Sim可以表示如下：

$$Sim(D_i,D_j)=\cos\theta=\frac{\sum_{k=1}^{n}W_k(D_i)\times W_k(D_j)}{\sqrt{\left(\left(\sum_{k=1}^{n}W_k^2(D_i)\right)\left(\sum_{k=1}^{n}W_k^2(D_j)\right)\right)}}$$

- $D_1 = 2K_1 + 3K_2 + 5K_3$
- $D_2 = 3K_1 + 7K_2 + K_3$
- $Q = 0K_1 + 0K_2 + 2K_3$



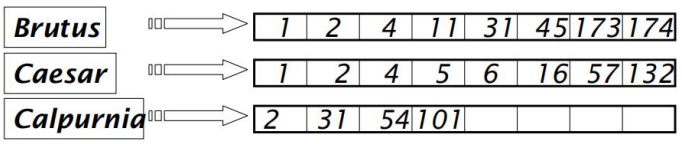

$$Sim(D_1,Q)=\frac{2\times0+3\times0+5\times2}{\sqrt{(2^2+3^2+5^2)\times(2^2)}}=\frac{5}{\sqrt{38}}=0.81$$

$$Sim(D_2,Q)=\frac{3\times0+7\times0+1\times2}{\sqrt{(3^2+7^2+1^2)\times(2^2)}}=\frac{1}{\sqrt{59}}=0.13$$

向量空间模型

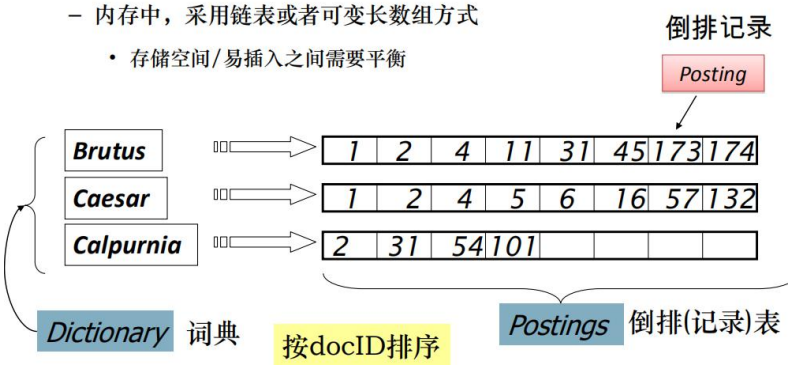
## 倒排索引 (Inverted index)

- 对每个词项t, 记录所有包含t的文档列表.
  - 每篇文档用一个唯一的 docID来表示，通常是正整数，如1,2,3…
- 能否采用定长数组的方式来存储docID列表

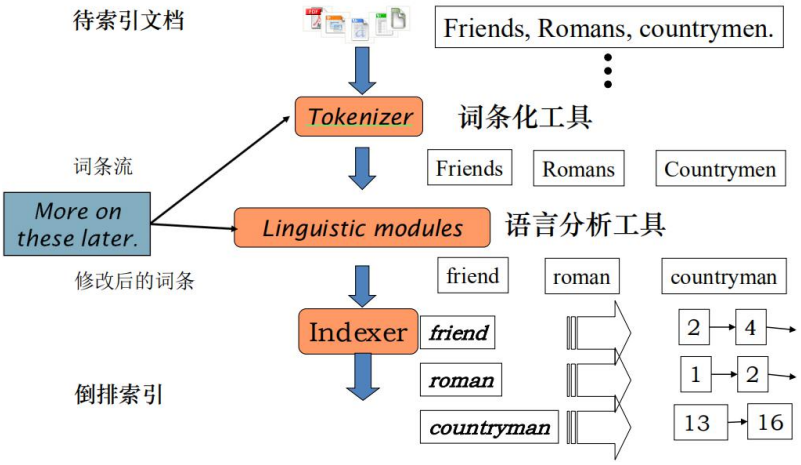



文档14中加入单词**Caesar**时该如何处理?

- 通常采用变长表方式
  - 磁盘上，顺序存储方式比较好，便于快速读取
  - 内存中，采用链表或者可变长数组方式
    - 存储空间/易插入之间需要平衡




## 倒排索引构建



# 索引构建过程: 词条序列

- <词条，docID>二元组

Doc 1

I did enact Julius Caesar I was killed i' the Capitol; Brutus killed me.

Doc 2

So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious

| Term | docID |
|---|---|
| I | 1 |
| did | 1 |
| enact | 1 |
| julius | 1 |
| caesar | 1 |
| I | 1 |
| was | 1 |
| killed | 1 |
| i' | 1 |
| the | 1 |
| capitol | 1 |
| brutus | 1 |
| killed | 1 |
| me | 1 |
| so | 2 |
| let | 2 |
| it | 2 |
| be | 2 |
| with | 2 |
| caesar | 2 |
| the | 2 |
| noble | 2 |
| brutus | 2 |
| hath | 2 |
| told | 2 |
| you | 2 |
| caesar | 2 |
| was | 2 |
| ambitious | 2 |

# 索引构建过程: 排序

- 按词项排序
  - 然后每个词项按docID**排序**

索引构建的核心步骤

| Term | docID |
|---|---|
| I | 1 |
| did | 1 |
| enact | 1 |
| julius | 1 |
| caesar | 1 |
| I | 1 |
| was | 1 |
| killed | 1 |
| i' | 1 |
| the | 1 |
| capitol | 1 |
| brutus | 1 |
| killed | 1 |
| me | 1 |
| so | 2 |
| let | 2 |
| it | 2 |
| be | 2 |
| with | 2 |
| caesar | 2 |
| the | 2 |
| noble | 2 |
| brutus | 2 |
| hath | 2 |
| told | 2 |
| you | 2 |
| caesar | 2 |
| was | 2 |
| ambitious | 2 |

| Term | docID |
|---|---|
| ambitious | 2 |
| be | 2 |
| brutus | 1 |
| brutus | 2 |
| capitol | 1 |
| caesar | 1 |
| caesar | 2 |
| caesar | 2 |
| did | 1 |
| enact | 1 |
| hath | 1 |
| I | 1 |
| I | 1 |
| i' | 1 |
| it | 2 |
| julius | 1 |
| killed | 1 |
| killed | 1 |
| let | 2 |
| me | 1 |
| noble | 2 |
| so | 2 |
| the | 1 |
| the | 2 |
| told | 2 |
| you | 2 |
| was | 1 |
| was | 2 |
| with | 2 |

# 索引构建过程: 词典 & 倒排记录表

- 某个词项在单篇文档中的多次出现会被合并
- 拆分成词典和倒排记录表两部分
- 每个词项出现的文档数目(doc. frequency, DF)会被加入

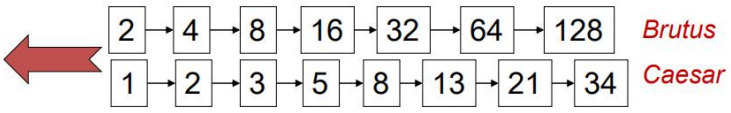
为什么加入DF值?

| Term | docID |
|---|---|
| ambitious | 2 |
| be | 2 |
| brutus | 1 |
| brutus | 2 |
| capitol | 1 |
| caesar | 1 |
| caesar | 2 |
| caesar | 2 |
| did | 1 |
| enact | 1 |
| hath | 1 |
| I | 1 |
| I | 1 |
| i' | 1 |
| it | 2 |
| julius | 1 |
| killed | 1 |
| killed | 1 |
| let | 2 |
| me | 1 |
| noble | 2 |
| so | 2 |
| the | 1 |
| the | 2 |
| told | 2 |
| you | 2 |
| was | 1 |
| was | 2 |
| with | 2 |

| term | doc. freq. | → | postings lists |
|---|---|---|---|
| ambitious | 1 | → | 2 |
| be | 1 | → | 2 |
| brutus | 2 | → | 1 → 2 |
| capitol | 1 | → | 1 |
| caesar | 2 | → | 1 → 2 |
| did | 1 | → | 1 |
| enact | 1 | → | 1 |
| hath | 1 | → | 2 |
| i | 1 | → | 1 |
| i' | 1 | → | 1 |
| it | 1 | → | 2 |
| julius | 1 | → | 1 |
| killed | 1 | → | 1 |
| let | 1 | → | 2 |
| me | 1 | → | 1 |
| noble | 1 | → | 2 |
| so | 1 | → | 2 |
| the | 2 | → | 1 → 2 |
| told | 1 | → | 2 |
| you | 1 | → | 2 |
| was | 2 | → | 1 → 2 |
| with | 1 | → | 2 |

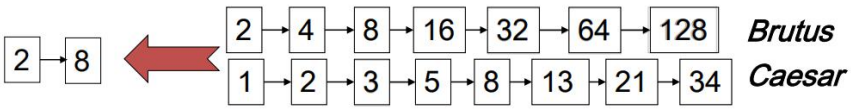
## 如何利用该索引来处理查询?

- AND 查询的处理
- 考虑如下查询（从简单的布尔表达式入手）：
  - Brutus AND Caesar
  - 在词典中定位 Brutus
    - 返回对应倒排记录表(对应的docID)
  - 在词典中定位Caesar
    - 再返回对应倒排记录表
  - 合并(Merge)两个倒排记录表，即**求交集**




## 合并过程

- 每个倒排记录表都有一个定位指针，两个指针同时从前往后扫描，每次比较当前指针对应倒排记录，然后移动某个或两个指针。合并时间为两个表长之和的线性时间.

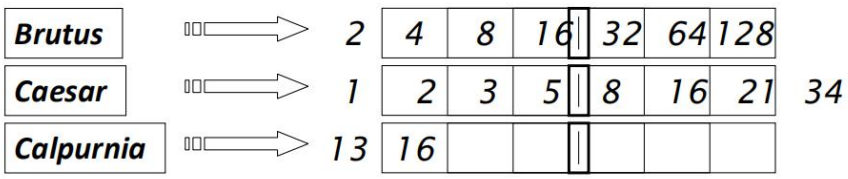



假定表长分别为x 和y, 那么上述合并算法的复杂度为 O(x+y)

关键原因: 倒排记录表按照docID排序

## 查询优化

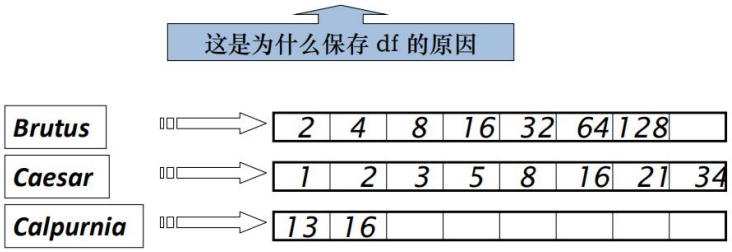
- 查询处理中是否存在处理的顺序问题?
- 考虑n 个词项的 AND
- 对每个词项，取出其倒排记录表，然后两两合并




查询: ***Brutus*** *AND* ***Calpurnia*** *AND* ***Caesar***

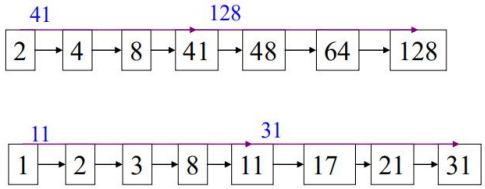
## 查询优化

- 按照表从小到大(即 df 从小到大)的顺序进行处理:
  - 每次从最小的开始合并




相当于处理查询 (*Calpurnia AND Brutus*) *AND Caesar*.

## 基于跳表指针的查询处理




假定匹配到上下的指针都指向8，接下来两个指针都向下移动一位。

比较**41和11，11小**

此时看11上面的跳表指针，指向31，31仍然比41小，于是下指针可以直接跳过中间的11、17、21、31

## tf，idf

列向量均值二值化，大于某个值为正，形成二值化矩阵
不会影响 tf 值，但是会影响 idf 值，因为采用的是出现与否的概念即所出现的文档数量为真的文档数量；就会变成行向量模型，看某个词有多少真，dtf 值为 3，idf 应该是 2；

## 词权： 词项频率 (term frequency，简称tf)

| | Anthony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth . . . |
|---|---|---|---|---|---|---|
| ANTHONY | 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 |
| . . . | | | | | | |

每篇文档可以看成是一个二值的向量 $\in \{0, 1\}^{|V|}$

# 非二值关联矩阵(词项频率)

| | Anthony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth . . . |
|---|---|---|---|---|---|---|
| ANTHONY | 157 | 73 | 0 | 0 | 0 | 1 |
| BRUTUS | 4 | 157 | 0 | 2 | 0 | 0 |
| CAESAR | 232 | 227 | 0 | 2 | 1 | 0 |
| CALPURNIA | 0 | 10 | 0 | 0 | 0 | 0 |
| CLEOPATRA | 57 | 0 | 0 | 0 | 0 | 0 |
| MERCY | 2 | 0 | 3 | 8 | 5 | 8 |
| WORSER | 2 | 0 | 1 | 1 | 1 | 5 |
| . . . | | | | | | |

每篇文档可以表示成一个词频向量 $\in N^{|V|}$

# 词权与词项频率

- 词袋(Bag of words)模型：BOW
- 词项t的词项频率(以下简称词频) $tf_{t,d}$ 是指t 在d中出现的次数，是与文档相关的一个量，可以认为是**文档内代表度**的一个量，也可以认为是一种**局部信息**。
- 第一种方法是采用原始的tf值(raw tf)
- 但是原始tf不太合适：
  - 某个词项在A文档中出现十次，即tf = 10，在B文档中 tf = 1，那么A比B更相关
  - 但是相关度不会相差10倍，即相关度不会正比于词项频率tf

# 一种替代原始tf的方法: 对数词频

- t 在 d 中的对数词频权重定义如下：

$$w_{t,d} = \begin{cases} 1 + \log_{10} tf_{t,d} & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$$

- $tf_{t,d} \rightarrow w_{t,d}$ : 0
→ 0, 1 → 1, 2 → 1.3, 10 → 2, 1000 → 4, 等等

- 文档-词项的匹配得分是所有同时出现在$q$和文档$d$中的词项的对数词频之和 $\sum_{t \in q \cap d} (1 + \log tf_{t,d})$
- 如果两者没有公共词项，则得分为0

## 罕见词的指示意义： idf 权重

- $df_t$ 是出现词项$t$的文档数目
- $df_t$ 是和词项$t$的信息量成反比的一个值
- 于是可以定义词项t的idf权重(逆文档频率):

$$idf_t = \log_{10} \frac{N}{df_t}$$

(其中$N$是文档集中文档的数目)

- $idf_t$ 是反映词项t的信息量的一个指标，是一种全局性指标，反应的是词项在全局的区别性。
- 实际中往往计算[log $N/df_t$ ]而不是 [$N/df_t$ ]，这可以对idf的影响有所抑制
- 值得注意的是，对于tf 和idf我们都采用了对数计算方式

## tf-idf权重计算

- 词项的tf-idf权重是tf权重和idf权重的乘积

$$w_{t,d} = (1 + \log tf_{t,d}) \cdot \log \frac{N}{df_t}$$

- 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 x idf

## TF 矩阵

| | Anthony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth . . . |
|---|---|---|---|---|---|---|
| ANTHONY | 157 | 73 | 0 | 0 | 0 | 1 |
| BRUTUS | 4 | 157 | 0 | 2 | 0 | 0 |
| CAESAR | 232 | 227 | 0 | 2 | 1 | 0 |
| CALPURNIA | 0 | 10 | 0 | 0 | 0 | 0 |
| CLEOPATRA | 57 | 0 | 0 | 0 | 0 | 0 |
| MERCY | 2 | 0 | 3 | 8 | 5 | 8 |
| WORSER | 2 | 0 | 1 | 1 | 1 | 5 |
| . . . | | | | | | |

每篇文档表示成一个词频向量$\in \mathbb{N}^{|V|}$

# 二值 → → tfidf矩阵

| | Anthony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth . . . |
|---|---|---|---|---|---|---|
| ANTHONY | 5.25 | 3.18 | 0.0 | 0.0 | 0.0 | 0.35 |
| BRUTUS | 1.21 | 6.10 | 0.0 | 1.0 | 0.0 | 0.0 |
| CAESAR | 8.59 | 2.54 | 0.0 | 1.51 | 0.25 | 0.0 |
| CALPURNIA | 0.0 | 1.54 | 0.0 | 0.0 | 0.0 | 0.0 |
| CLEOPATRA | 2.85 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| MERCY | 1.51 | 0.0 | 1.90 | 0.12 | 5.25 | 0.88 |
| WORSER | 1.37 | 0.0 | 0.11 | 4.15 | 0.25 | 1.95 |
| . . . | | | | | | |

每篇文档表示成一个基于tfidf权重的实值向量 $\in R^{|V|}$

# 向量空间模型

- 标引词的权重计算（TF-IDF）
  - N为文档集合，$n_i$为包含标引词$K_i$的文档篇数，$TF_{ij}$表示标引词$K_i$在文档$D_j$中出现的频数，则文档$D_j$中标引词$K_i$的标准化频率$F_{ij}$为

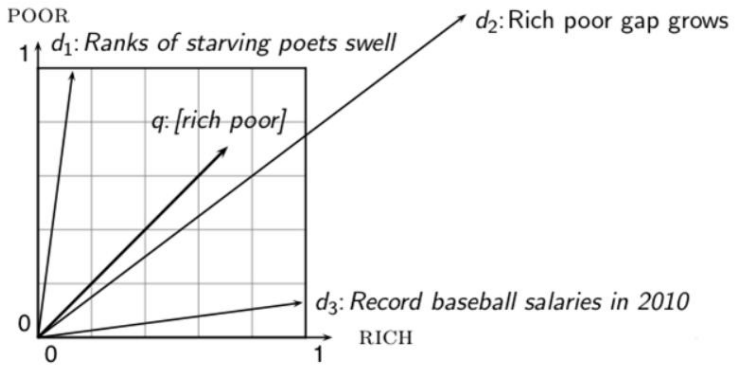
$$F_{ij} = TF_{ij} / \max_j TF_{ij}$$

  - 最大值是通过计算文档$D_j$中出现的所有标引词来获得的。如果标引词$K_i$没有出现在文档$D_j$中，则$F_{ij} = 0$。
  - 标引词$K_i$的IDF为$IDF_i = \log(N/n_i)$
  - 标引词$K_i$在文档$D_j$中的权重$W_{ij} = F_{ij} * IDF_i$

# 查询看成向量

- 关键思路1: 对于查询做同样的处理，即将查询表示成同一高维空间的向量
- 关键思路2: 按照文档对查询的邻近程度排序
  - 邻近度 = 相似度
  - 邻近度≈ 距离的反面
- 回想一下，我们是希望和布尔模型不同，能够得到非二值的、既不是过多或也不是过少的检索结果
- 通过计算出相关文档的相关度高于不相关文档相关度的方法来实现

# 欧氏距离 vs 余弦距离

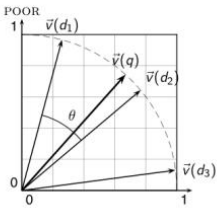



尽管查询$q$和文档$\vec{d}_2$的词项分布非常相似，但是采用欧氏距离计算它们对应向量之间的距离非常大。

夹角相似：在区间[0◦, 180◦]上，余弦函数cosine是一个单调递减函数

# 查询和文档之间的余弦相似度计算

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




- $q_i$ 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
- $|\vec{q}|$ 和 $|\vec{d}|$ 分别是 $\vec{q}$ 和 $\vec{d}$ 的长度
- 上述公式就是 $\vec{q}$ 和 $\vec{d}$ 的余弦相似度，或者说向量 $\vec{q}$ 和 $\vec{d}$ 的夹角的余弦

# 余弦相似度的计算样例

3本小说之间的相似度

(1) SaS(理智与情感):

Sense and Sensibility

(2) PaP(傲慢与偏见):

Pride and Prejudice

(3) WH(呼啸山庄):

Wuthering Heights

词项频率tf

| 词项 | SaS | PaP | WH |
|---|---|---|---|
| AFFECTION | 115 | 58 | 20 |
| JEALOUS | 10 | 7 | 11 |
| GOSSIP | 2 | 0 | 6 |
| WUTHERING | 0 | 0 | 38 |

# 余弦相似度计算

词项频率 tf

| 词项 | SaS | PaP | WH |
|---|---|---|---|
| AFFECTION | 115 | 58 | 20 |
| JEALOUS | 10 | 7 | 11 |
| GOSSIP | 2 | 0 | 6 |
| WUTHERING | 0 | 0 | 38 |

对数词频（$1+\log_{10}tf$）

| 词项 | SaS | PaP | WH |
|---|---|---|---|
| AFFECTION | 3.06 | 2.76 | 2.30 |
| JEALOUS | 2.0 | 1.85 | 2.04 |
| GOSSIP | 1.30 | 0 | 1.78 |
| WUTHERING | 0 | 0 | 2.58 |

为了简化计算，上述计算过程中没有引入IDF

# 余弦相似度计算

对数词频$(1+\log_{10}tf)$

| 词项 | SaS | PaP | WH |
|---|---|---|---|
| AFFECTION | 3.06 | 2.76 | 2.30 |
| JEALOUS | 2.0 | 1.85 | 2.04 |
| GOSSIP | 1.30 | 0 | 1.78 |
| WUTHERING | 0 | 0 | 2.58 |

对数词频的余弦归一化结果

| 词项 | SaS | PaP | WH |
|---|---|---|---|
| AFFECTION | 0.789 | 0.832 | 0.524 |
| JEALOUS | 0.515 | 0.555 | 0.465 |
| GOSSIP | 0.335 | 0.0 | 0.405 |
| WUTHERING | 0.0 | 0.0 | 0.588 |

$\cos(SaS, PaP) \approx 0.789 * 0.832 + 0.515 * 0.555 + 0.335 * 0.0 + 0.0 * 0.0 \approx 0.94.$

$\cos(SaS, WH) \approx 0.79$

$\cos(PaP, WH) \approx 0.69$

$\cos(SaS, PaP) > \cos(SAS, WH) > \cos(PaP, WH)$