



第2章 万维网网页信息的表达及解析

2.1 万维网架构及网页表达

2.2 网页信息抽取

2.3 搜索引擎与分析系统

2.4 搜索引擎索引系统

2.5 搜索引擎查询系统5

《走进搜索引擎》 潘雪峰 花贵春 梁斌编著
电子工业出版社 2011年5月第2版



2.5 搜索引擎查询系统

2.5.1 相关概念

2.5.2 网页信息检索

- 1.信息检索模型
- 2.布尔模型
- 3.向量空间模型
- 4.关键词权重的量化方法TF/IDF

2.5.3 搜索引擎检索与生成结果页

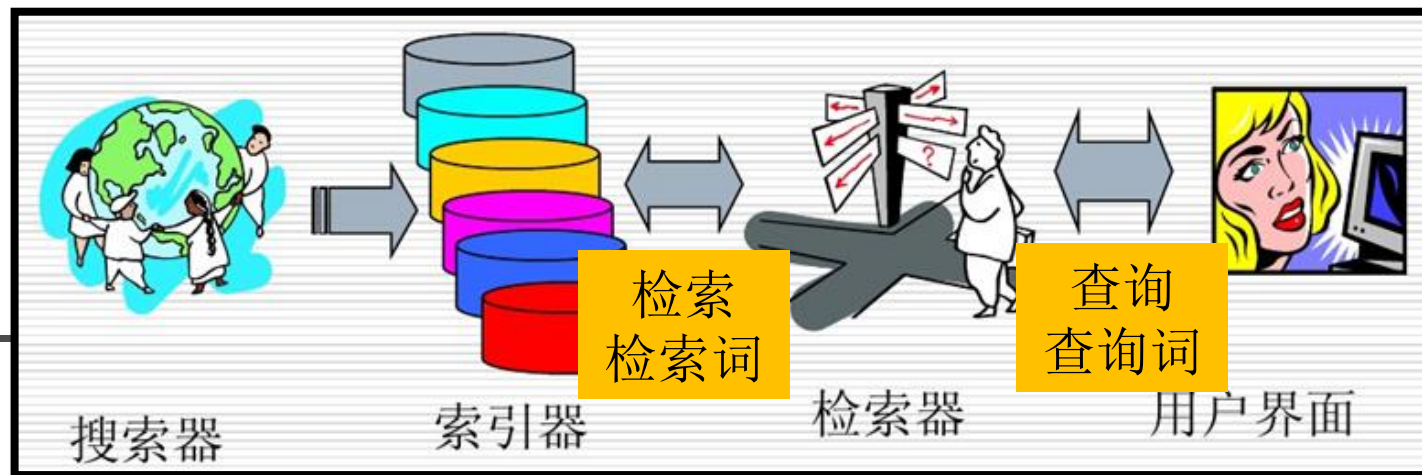


2.5.1 相关概念

■ 搜索引擎查询系统——网页信息检索

- 负责分析用户提交的查询请求
- 进行检索、排序及自动文本摘要提取
- 以文本摘要作为搜索结果返回给用户

- 从文档中自动提取出的一个正文片断。
- 用户仅仅需要浏览整个正文片段就能够了解文档中与查询词相关的部分，进而判断是否值得详细阅读整篇文档



■ 查询

- 真实用户进行的一次查询，是相对于搜索引擎查询系统而言的；
- 提交给查询系的关键词称为“查询词”；
- 结果是搜索结果网页

■ 检索

- 检索代理对索引库进行的一次检索，是相对于搜索引擎索引系统而言的
- 提交检索代理的称为“检索词”
- 结果是与查询词相关的文档列表

• 用户提交查询词为“清华大学图书馆”

简化:统一使用查询词

• 通过分词，提交给检索代理变成“清华大学”和“图书馆”两个检索词。



2.5 搜索引擎查询系统

2.5.1 相关概念

2.5.2 网页信息检索

1.信息检索模型

2.布尔模型

3.向量空间模型

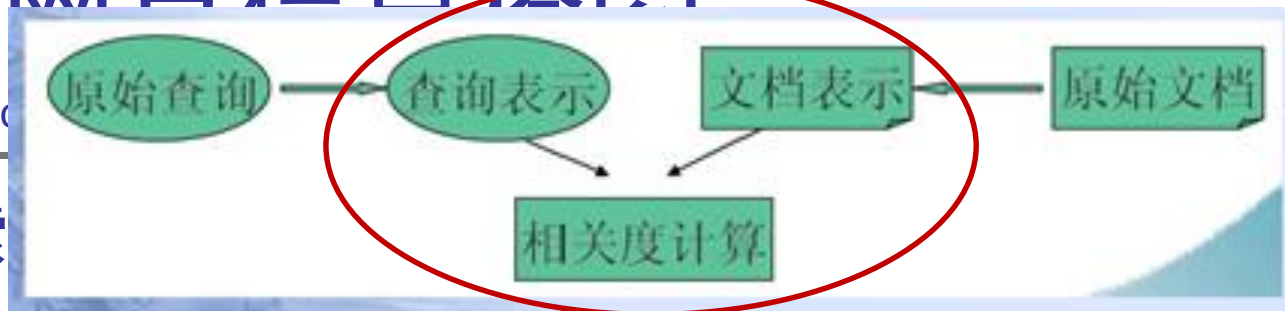
4.关键词权重的量化方法TF/IDF

2.5.3 搜索引擎检索与生成结果页

2.5.2

网络信息检索

<https://wenku.baidu.com>



■ 1.信息检索

■ 信息检索

- 出现于20世纪50年代
- 指从信息资源的集合中查找所需文献或查找所需文献中包含的信息内容的过程。

■ 信息检索模型

- 如何对查询和文档进行表示
- 如何进行相似度计算的框架和方法，本质上是对相关度建模
- 信息检索模型是信息检索的核心内容之一。



2.5.2 网页信息检索

<https://wenku.baidu.com/view/da6a996886c24028915f804d2b160b4e777f8113.html>

■ 网页信息检索

- 使用搜索引擎完成的特殊信息检索
- 数据源来自于网页索引库

■ 网页信息检索模型

- 布尔模型——早期
- 向量空间模型
- 关键词权重的量化方法TF/IDF



2.5.2 网页信息检索

2. “布尔模型” (Boolean Models)

- 早期的检索模型，也称为“集合模型”
 - 采用AND、OR及NOT等逻辑运算符将多个查询词连成一个逻辑表达式
 - 通过布尔运算进行检索的简单匹配模型。
- 例：查询词为“走进搜索引擎 检索模型—搜索”，翻译成“走进搜索引擎AND检索模型NOT搜索”

应用最广泛的模型；
目前仍然应用于商业系统中；

布尔模型(Boolean Model)

CSDN博主「iterate7」的原创文章,

<https://blog.csdn.net/iterate7/article/details/77206613>

■ 布尔模型描述:

■ 文档D:

- 一个文档被表示为关键词的集合

■ 查询式Q:

- 被表示为关键词的布尔组合, 用“与、或、非”连接起来, 并用括弧指示优先次序

■ 匹配F:

- 一个文档当且仅当它能够满足布尔查询式时, 才将其检索出来

■ 检索策略

- 基于二值判定标准

布尔模型检索示例

■ 示例

文档集包含两个文档:

文档1: a b c f g h

文档2: a f b x y z

查询向量 (a, b, c)

文档1: (1, 1, 1)

文档2: (1, 1, 0)

用户查询: 文档中出现a或者b, 但一定要出现z。

将查询表示为布尔表达式,并转换成析取范式

查询: (a OR b) AND c

返回文档1

2.5.2 网页信息检索

- 查询2006年（当年）超女5进4比赛的新闻，用布尔模型怎么构造查询？

(2006 OR 今年) AND (超级女声 OR 超女 OR 超级女生) AND (6进5 OR 六进五 OR 六AND 进AND 五)

表达式相当复杂，构造困难！

不严格的话结果过多，而且很多不相关；
非常严格的话结果会很少，漏掉很多结果。

- 文档：

D1:...据报道计算机病毒最近猖獗 ✓

D2: 小王虽然是学医的，但对研究电脑病毒也感兴趣... ✗

D3: 计算机程序发现了艾滋病病毒传播途径 ✓

- $Q = \text{病毒} \text{AND} (\text{计算机} \text{OR} \text{电脑}) \text{ANDNOT} \text{医}$

■ 示例：查找搜索引擎系统构成

查询“搜索引擎 系统构成”

- 布尔模型检索结果：
 - 文档1和文档2被检索到
 - 文档3没捡到

(1) 在传统搜索引擎架构中，搜索引擎由4个系统构成，分别是下载系统、分析系统、索引系统及查询系统。



(2) 机械行业内一般把小型挖掘简称为“小挖”，小挖由5个系统构成，分别是.....，详细地理解这些名词可以使用Google搜索引擎搜索一下。



(3) 搜索引擎有4个主要功能模块，分别是下载系统，分析系统，索引系统和查询系统。这4个系统是搜索引擎的核心，其中查询系统是搜索引擎唯一直接面对客户的系统。



从用户查询意图上看，文档3比文档2更加符合用户的查询意图，仅仅因为没有包含“系统构成”这个关键词，而没有被检索出——二值判断的弊端



2.5.2 网页信息检索

■ “布尔模型”

■ 优点：

- 易于实现，检索速度快。

■ 缺点：

- 没有考虑文档和查询词的相关性问题
- 没有区分查询词的权重问题，“效果”差；
- 检索结果无法排序

- 布尔模型中很难进行相关性强弱的度量
- 只解决“有”还是“没有”的问题
- 不解决“好”还是“不好”的问题。

2.5.2 网页信息检索

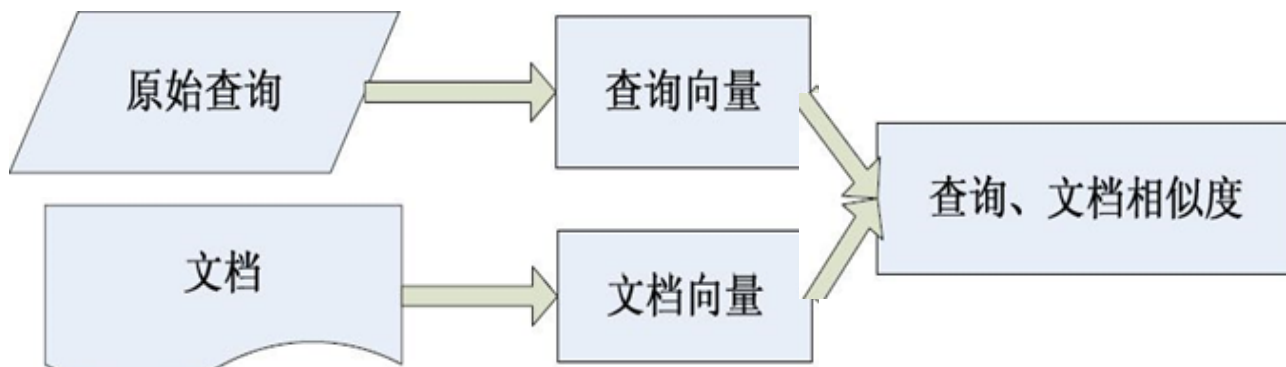
3. 向量空间模型 (VSM: Vector Space Model)

■ 历史

- Salton在上世纪60年代提出的特征表达系统理论框架
- 成功应用于SMART文本检索系统;
- 现在仍然是信息检索技术研究的基础。

■ VSM模型描述

- 文档被表示为文档空间的向量, 通过计算向量之间的相似性来度量文档间的相似性。
- 文本处理中最常用的相似性度量方式是余弦距离。

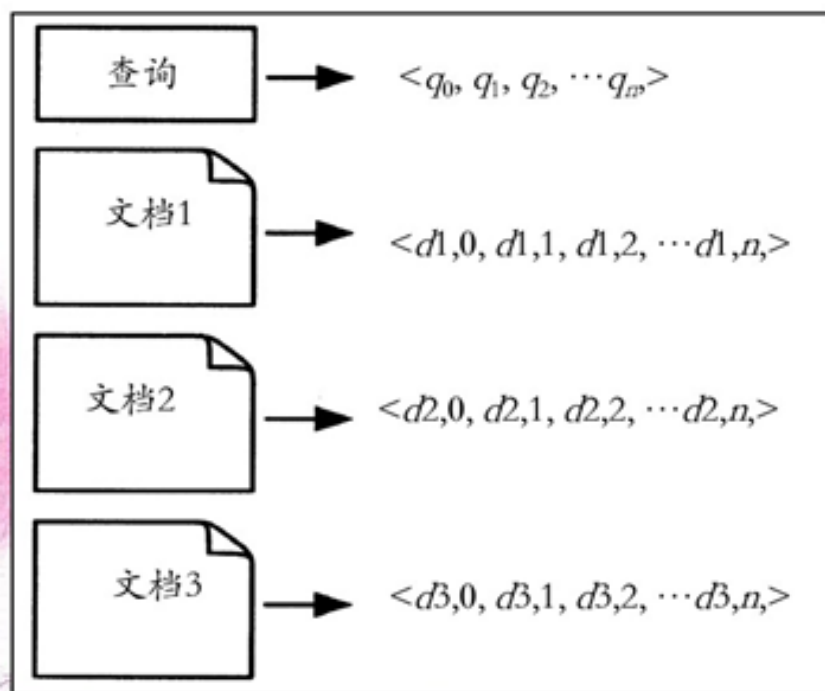


2.5.2 网页信息检索

<https://wenku.baidu.com/view/1f59dc7d4a35eefdc8d376eeaeaad1f3469311b1.html>

■ 向量空间模型主要工作：

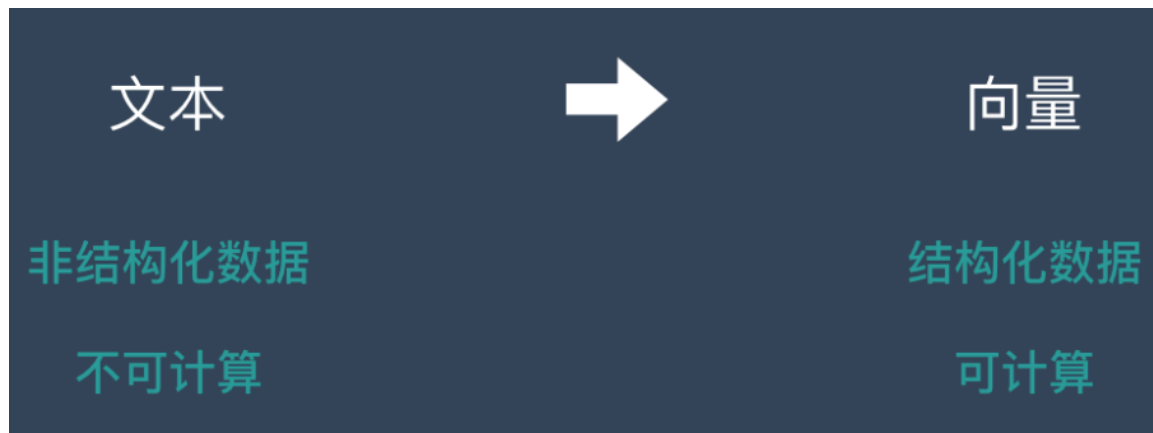
- 将查询词和文档按照关键词的维度分别向量化
- 计算这两个向量间相似度，即文档与查询词的相似度
- 优先检索与查询词相似度大的文档



- 余弦相似度
- 皮尔森相关系数
- Jaccard相似系数
- 对数似然相似率
- 信息检索-词频-逆文档频率 (TF-IDF)
-

2.5.2 网页信息检索

- 向量（vector）：
 - 又称为“矢量”。最初被应用于物理学，很多物理量，如力、速度、位移、电场强度等
- 向量包含了两层含义：
 - 长度：用向量的模表示，每个分量的平方和开根号
 - 方向：被用来量化向量的相似程度。
- 文本向量化可以将一些文本处理问题转换为向量计算问题。



2.5.2 网页信息检索

■ 文本特征向量化方法：

1.词集模型：one-hot编码向量化文本（统计各词在文本中是否出现） <https://blog.csdn.net/shanhui123/article/details/99440278>

- 将处理的文本用二进制进行表示，每个字的维度是字典的大小
- 例：处理文本是：“自然语言处理”
- 该字典有6个字，“自然语言处理”，维度为6。出现的标1，否则标0.

自：[1 0 0 0 0 0]

然：[0 1 0 0 0 0]

语：[0 0 1 0 0 0]

言：[0 0 0 1 0 0]

处：[0 0 0 0 1 0]

理：[0 0 0 0 0 1]

- 维度大，稀疏
- 词集模型没有考虑词的频率

2.5.2 网页信息检索

■ 文本特征向量化方法不断改进：

1.词集模型：词集模型没有考虑词的频率

2.词袋模型：文档中出现的词对应的one-hot向量相加（统计各词在文本中出现次数，在词集模型的基础上。）

词袋模型没有考虑词的重要度

3.词袋模型+IDF：TFIDF向量化文本（词袋模型+IDF值，考虑了词的重要性）

词袋模型+IDF没有考虑词的顺序

4.N-gram模型：考虑了词的顺序

N-gram模型的优点是考虑了词的顺序，但是会出现词表膨胀的问题

5.word2vec模型：使用文章中所有词的平均词向量作为文章的向量

- 例：处理文本“走进搜索引擎，学习搜索引擎”
- 词典：“走进”、“搜索引擎”和“学习” 3个词
- 词组成的向量空间就是我们熟悉的三维空间

走进

学习

在三维向量空间：

（“搜索引擎”，“走进”，“学习”）
向量化的结果：（2，1，1）
（采用词频做向量值）

走进搜索引擎，学习搜索引擎
（2，1，1）

一般汉语词汇大约5000条，如果用这5000维的词向量空间表示这个句子，向量十分稀疏

搜索引擎

假定向量空间为四维：

（“搜索引擎”，“走进”，“学习”，“检索模型”）
向量化的结果：（2，1，1，0）



2.5.2 网页信息检索

- 关键词的出现次数归一化
 - 为保证对大文档和小文档做到公平，对关键词的出现次数转化为词频（词数 / 总词数）作为向量的分量
- 上例的向量空间为四维：
- （“搜索引擎”，“走进”，“学习”，“检索模型”）向量化的结果：
(2, 1, 1, 0)

向量归一化 (2/4, 1/4, 1/4, 0)



2.5.2 网页信息检索

- 计算文档和查询词相似度的问题

- 方法1：向量之间的夹角余弦值判断向量相似度
- 其它方法：欧氏距离、Jaccard距离、编辑距离等

- 向量间的夹角余弦的计算公式：

$$\cos \theta = \frac{a \bullet b}{|a| \times |b|}$$

- 其中：

- a ， b 表示向量
- \bullet 表示向量的点乘
- $|a|$ 、 $|b|$ 表示向量的模，或者说是向量的长度

2.5.2 网页信息检索

例.假定在一个7维的向量空间下

- 一个查询词向量化为 \mathbf{a} (0, 0, 2, 0, 1, 0, 1)
- 一个文档向量化为 \mathbf{b} (0, 1, 3, 5, 2, 4, 0)
- 夹角余弦计算方法如下:

$$\mathbf{a} \bullet \mathbf{b} = (0, 0, 2, 0, 1, 0, 1) \cdot (0, 1, 3, 5, 2, 4, 0)^T$$

$$= 0 \cdot 0 + 0 \cdot 1 + 2 \cdot 3 + 0 \cdot 5 + 1 \cdot 2 + 0 \cdot 4 + 1 \cdot 0$$

$$= 8$$

$$\cos \theta = \frac{\mathbf{a} \bullet \mathbf{b}}{|\mathbf{a}| \times |\mathbf{b}|} = \frac{8}{2.45 \cdot 7.42} = 0.44$$

$$|\mathbf{a}| = \sqrt{0^2 + 0^2 + 2^2 + 0^2 + 1^2 + 0^2 + 1^2} = \sqrt{6} = 2.45$$

$$|\mathbf{b}| = \sqrt{0^2 + 1^2 + 3^2 + 5^2 + 2^2 + 4^2 + 0^2} = \sqrt{55} = 7.42$$

- 查询词 \mathbf{a} 和文档 \mathbf{b} 的相关性就转化为0.44
- 相似性量化的结果称为“相似度”

例.找相似文章(用余弦相似性)

$$\cos \theta = \frac{a \bullet b}{|a| \times |b|}$$

- 句子A: 我喜欢看电视, 不喜欢看电影
- 句子B: 我不喜欢看电视, 也不喜欢看电影
- 基本思路是: 如果这两句话的用词越相似, 它们的内容就应该越相似

1、分词

- 句子A: 我/喜欢/看/电视, 不/喜欢/看/电影。
- 句子B: 我/不/喜欢/看/电视, 也/不/喜欢/看/电影。

2、列出所有值

- 我, 喜欢, 看, 电视, 电影, 不, 也。

词典向量维度7

3、计算词频

- 句子A: 我 1, 喜欢 2, 看 2, 电视 1, 电影 1, 不 1, 也 0。

句子A词频向量: [1, 2, 2, 1, 1, 1, 0]

- 句子B: 我 1, 喜欢 2, 看 2, 电视 1, 电影 1, 不 2, 也 1

句子B词频向量: [1, 2, 2, 1, 1, 2, 1]

2.5.2 网页信息检索

- 在实际计算中，如果向量 \mathbf{a} 表示查询向量， \mathbf{a} 总是不变的
- 因此相似度计算简化为：

$$\cos \theta = \frac{\mathbf{a} \bullet \mathbf{b}}{|\mathbf{a}| \times |\mathbf{b}|} \quad \rightarrow \quad |\mathbf{a}| \times \cos \theta = \frac{\mathbf{a} \bullet \mathbf{b}}{|\mathbf{b}|}$$



4. 关键词权重的量化方法TF/IDF

- 向量空间模型以计数为特征，无法反映词汇重要程度
- 关键词权重的量化方法TF/IDF
 - 是一种用于信息检索的常用加权技术。
 - 是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度
- TF/IDF思想：
 - 字词的重要性随着它在**文件中**出现的次数成正比增加
 - 同时会随着它在**语料库**中出现的频率成反比下降。

TF/IDF含义：一个词语在一篇文章中出现次数越多，同时所有文档中出现次数越少，越能够代表该文章。



4. 关键词权重的量化方法TF/IDF

■ 词频 (term frequency, TF)

- 一个给定的词语在该文件中出现的次数。
- 这个数字通常会被归一化(一般是词频除以文章总词数), 以防止它偏向长的文件。

$$\text{词频(TF)} = \frac{\text{某个词在文章中的出现次数}}{\text{文章的总词数}}$$

<https://blog.csdn.net/zhaomengszu>

$$\text{词频(TF)} = \frac{\text{某个词在文章中的出现次数}}{\text{该文出现次数最多的词的出现次数}}$$

稀疏减少

4. 关键词权重的量化方法TF/IDF

<https://blog.csdn.net/zhaomengszu/article/details/81452907>

■ 仅有词频的不足：

- 一个词在文章中出现很多次，不一定有着很大的作用
 - 如：我们在一篇文档中，可能发现"中国"、"蜜蜂"、"养殖"这三个词的出现次数一样多
 - 因为"中国"是很常见的词，相对而言，"蜜蜂"和"养殖"不那么常见。如果这三个词在一篇文章的出现次数一样多，有理由认为，"蜜蜂"和"养殖"的重要程度要大于"中国"，也就是说，在关键词排序上面，"蜜蜂"和"养殖"应该排在"中国"的前面。

■ 解决方法：

- 停用词：去掉没有用的词：‘的’，‘是’这样的词
- 重要性调整系数：衡量一个词是不是常见词
 - 如果某个词比较少见，但是它在这篇文章中多次出现，那么它很可能就反映了这篇文章的特性，正是我们所需要的关键词

4. 关键词权重的量化方法TF/IDF

<https://blog.csdn.net/zhaomengszu/article/details/81452907>

- IDF: 逆文档频率(inverse document frequency)
 - 一个词语普遍重要性的度量
 - 在词频的基础上, 对每个词分配一个"重要性"权重

$$\text{逆文档频率(IDF)} = \log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数} + 1}\right)$$

- 语料库: 用来模拟语言的使用环境
- 如果一个词越常见, 那么分母就越大, 逆文档频率就越小
- 分母加1, 避免分母为0 (即所有文档都不包含该词)

4. 关键词权重的量化方法TF/IDF

<https://blog.csdn.net/zhaomengszu/article/details/81452907>

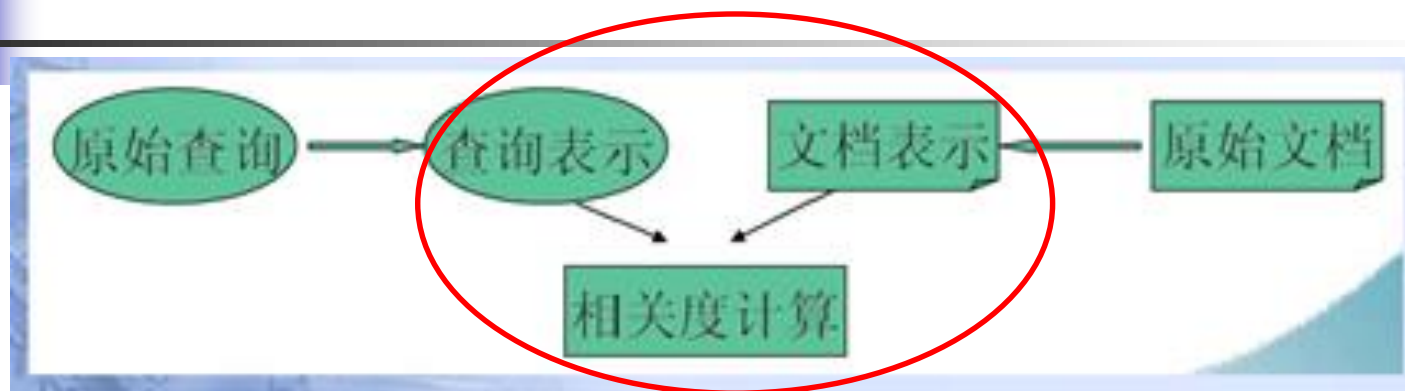
■ TF-IDF值：

- "词频" (TF) 和"逆文档频率" (IDF) 相乘
- 某个词对文章的重要性越高， TF-IDF值就越大
- 排在最前面的几个词，就是这篇文章的关键词。


$$\text{TF-IDF} = \text{词频(TF)} \times \text{逆文档频率 (IDF)}$$

- TF-IDF与一个词在文档中的出现次数成正比，与该词在整个语言中的出现次数成反比。

小结



	文档表示	查询表示	匹配	性能
布尔模型	关键词集合	关键词 布尔组合	二值判定	速度快 效果差
向量空间模型	向量 (TF)	向量	向量余弦	效果较上好 速度慢
TF-IDF方法	向量 (TF-IDF)	向量	比较 TF-IDF 值	效果更好 速度较快

- 
- 例：提取关键词
 - 一篇文章《中国的蜜蜂养殖》，假定该文长度为1000个词，"中国"、"蜜蜂"、"养殖"各出现20次
 - 则这三个词的"词频"（TF）都为0.02。
 - 然后，搜索Google发现：
 - 包含"的"字的网页共有250亿张，假定为中文网页总数
 - 包含"中国"的网页共有62.3亿张，
 - 包含"蜜蜂"的网页为0.484亿张，
 - 包含"养殖"的网页为0.973亿张
 - TF-IDF值的大小与一个词的常见程度成反比
 - 最常见的词（"的"、"是"、"在"）给予最小的权重
 - 较常见的词（"中国"）给予较小的权重
 - 较少见的词（"蜜蜂"、"养殖"）给予较大的权重

- 文章《中国的蜜蜂养殖》"中国"、"蜜蜂"、"养殖"的逆文档频率（IDF）和TF-IDF如下：

	包含该词的文档数（亿）	IDF	TF-IDF
中国	62.3	0.603	0.0121
蜜蜂	0.484	2.713	0.0543
养殖	0.973	2.410	0.0482

- "蜜蜂"的TF-IDF值最高，"养殖"其次，"中国"最低。
- 如果只选择一个词，"蜜蜂"就是这篇文章的关键词



4. 关键词权重的量化方法TF/IDF

■ TF-IDF算法用途：

- 自动提取关键词
- 信息检索

■ TF-IDF算法的优点

- 简单快速，结果比较符合实际情况。

■ 缺点

- 算法无法体现词的位置信息，出现位置靠前的词与出现位置靠后的词，都被视为重要性相同，这是不正确的。



2.5 搜索引擎查询系统

2.5.1 相关概念

2.5.2 网页信息检索

1.信息检索模型

2.布尔模型

3.向量空间模型

4.关键词权重的量化方法TF/IDF

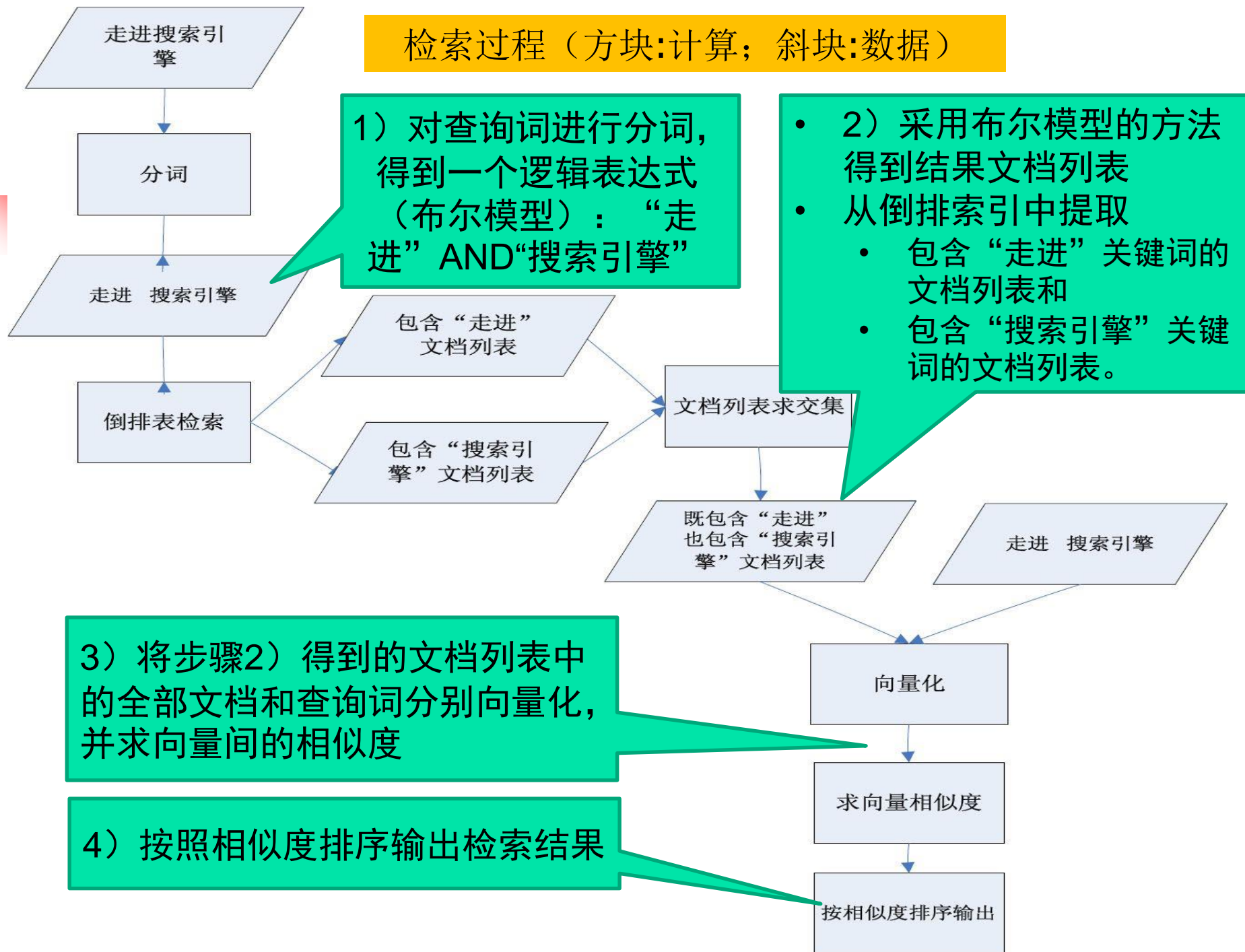
2.5.3 搜索引擎检索与生成结果页



2.5.3 搜索引擎检索与生成结果页

- 搜索引擎采用的检索模型
 - “布尔模型和向量空间模型”结合的方法来进行信息检索
 - 布尔模型的检索效率高且易于实现
 - 向量空间模型能够提高检索的相似度，通过相似度排序的手段能够大大改善查询效果

检索过程（方块:计算；斜块:数据）





2.5.3 搜索引擎检索与生成结果页

■ 详细的检索过程:

- (1) 对查询词进行分词，得到一个逻辑表达式。例如查询“走进搜索引擎”，将会被切分成“走进”，“搜索引擎”这两个词。并且转换为用**AND**逻辑表示的表达式，即“走进” **AND** “搜索引擎”。
- (2) 采用布尔模型的方法得到结果文档列表，例如从倒排索引中提取包含“走进”关键词的文档列表和包含“搜索引擎”关键词的文档列表。并将检索出的文档列表求交集，得到既包含“走进”，也包含“搜索引擎”的文档列表。
- (3) 将步骤(2)得到的文档列表中的全部文档和查询词分别向量化，并求向量间的相似度。
- (4) 按照相似度排序输出检索结果。

生成搜索结果页全过程

