小组成员

1	2	3	4
			
2120101740	2120101750	2120101755	2120101759
李东艳	马静	农倩倩	邱秀珍

课程布局

一 布尔检索、邻近检索以及相关度计算

二 使用关系模式进行半结构化搜索

三 引入vpn虚拟专用网络相关背景

四 介绍vpn虚拟专用网络及其安全隐患

信息检索作为关系应用

李东艳

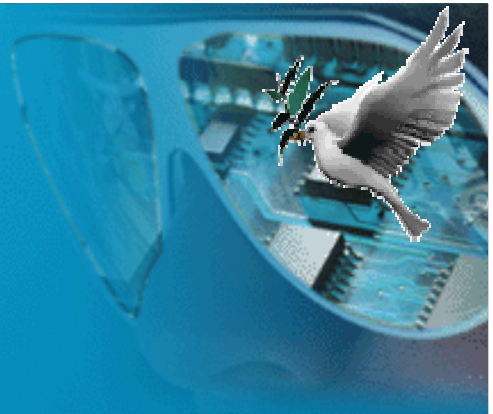
指导老师：赵燕平　张华平

2010年12月7日

LOGO

如何从文献的迷宫中获取我需要的信息?


Google 高级搜索 - 360安全浏览器 3.5 正式版
文件(F) 查看(V) 收藏(B) 帐户(U) 工具(T) 帮助(H)
后退 前进 停止 刷新 主页 恢复 收藏 历史 无痕
http
allintext: 布尔检
登录 添加 谷歌 IE打开 充值 翻译 截图 实用工具
Google 高级搜索
搜索帮助 | Google 大全
搜索结果
包含全部字词
布尔检索
10 项结果
包含完整字句
Google 搜索
包含至少一个字词
不包括字词
语言
搜索网页语言是
简体中文
地区
搜索网页位置于:
任何国家/地区
文件格式
仅
显示使用以下文件格式的结果
Microsoft Powerpoint (.ppt)
日期
返回在下列时间段内首次查看的网页
一年内
字词位置
查询字词位于
网页的正文
网站
仅
搜索以下网站
完毕
下载
100%



第一部分　布尔检索

1. 布尔模型的基本原理

两条基本规则：

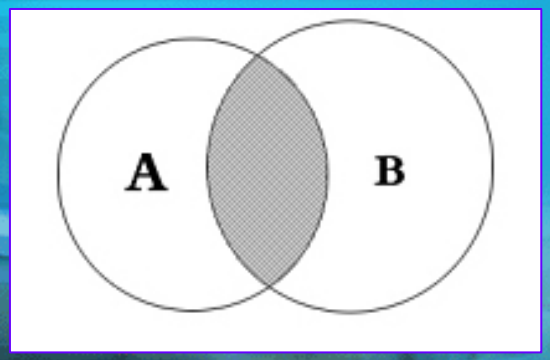
① 系统索引词集合中的每一个索引词在一篇文档中只有两种状态：出现或者不出现。相应地，每个索引词的权值 wij ∈{0，1}；

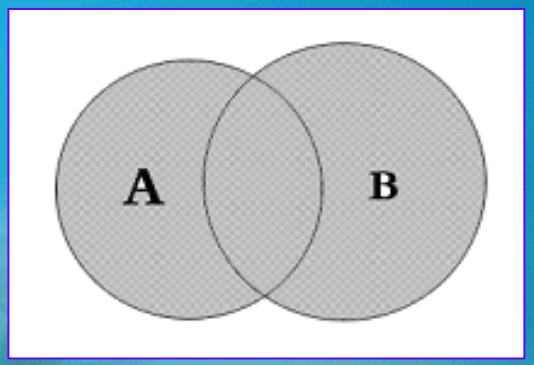
② 用户提问式q由三种布尔逻辑运算符“and”、“or”、“not”连接索引词构成。

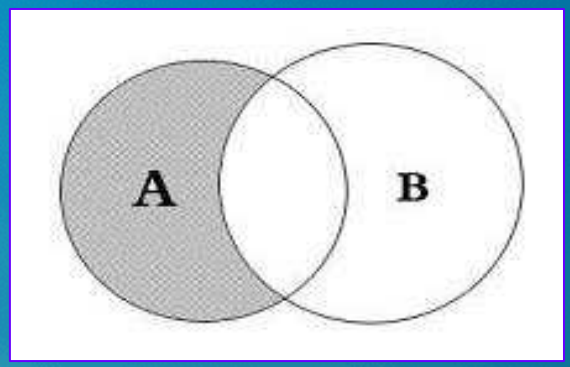
- 文档表示
  - 一个文档被表示为关键词的集合
- 查询式表示
  - 查询式(Queries)被表示为关键词的布尔组合，用“与、或、非”连接起来，并用括弧指示优先次序。
- 匹配
  - 一个文档当且仅当它能够满足布尔查询式时，才将其检索出来。

布尔操作(关系):

- 与(AND)： A AND B = true if A=B=true
- 或(OR) ： A OR B = true if A=true OR B=true
- 非(NOT) ： NOT A = true if A=false

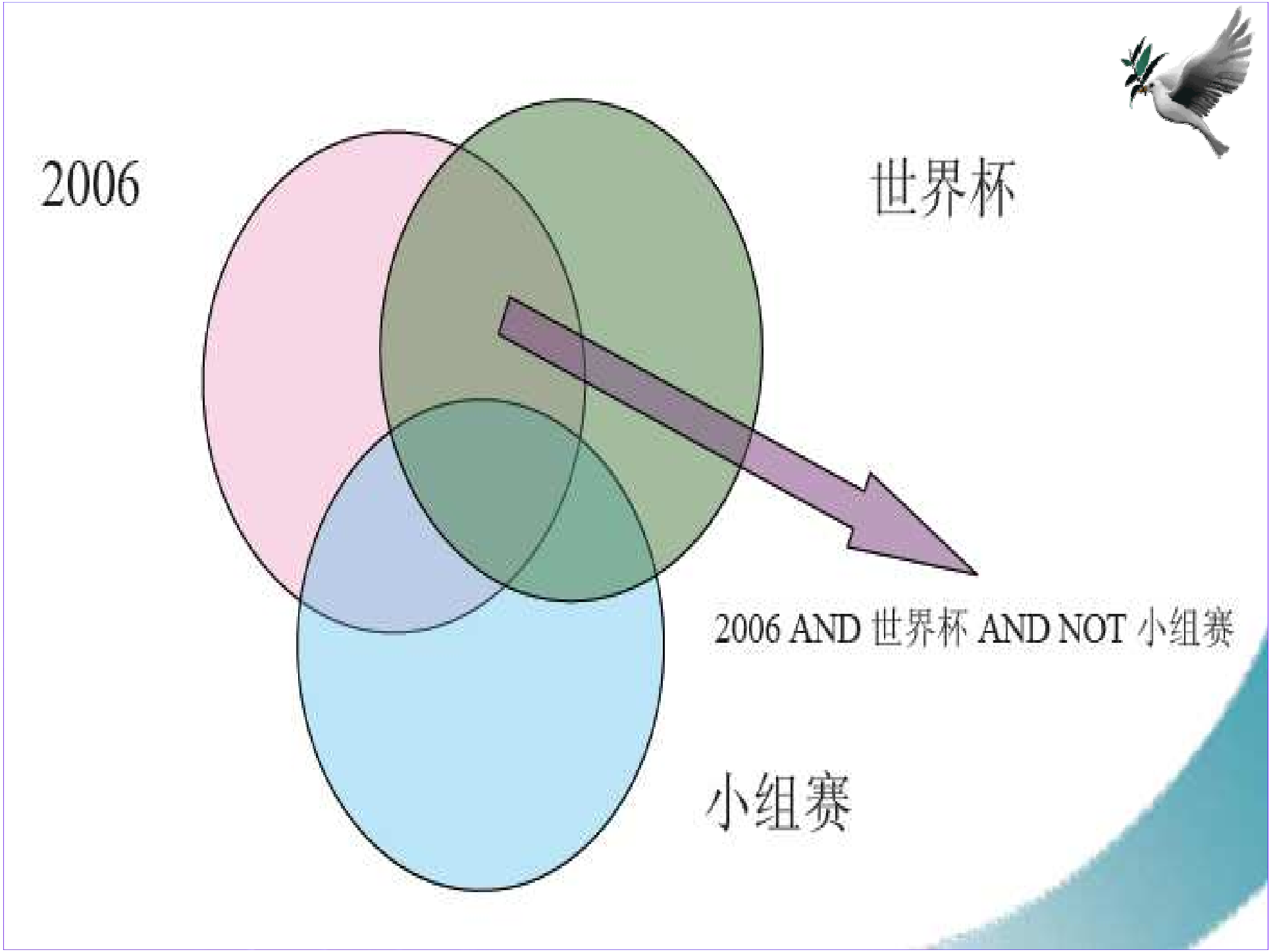






实例一

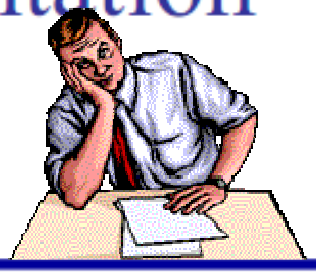
- 查询：2006 AND 世界杯 AND NOT 小组赛
- 文档1：2006年世界杯在德国举行。
- 文档2：2006年世界杯小组赛已经结束。
- 相似度计算：布尔查询表达式和所有文档的布尔表达式进行匹配，匹配成功得分为1，否则为0。


2006
世界杯
2006 AND 世界杯 AND NOT 小组赛
小组赛

LOGO

2. 使用SQL实现布尔检索

Relational Document Representation



DOCUMENT

docID	docname	headline	dateline
28	AP881214-0028	Stocks Up In Tokyo	TOKYO (AP)

INDEX

docID	termcnt	term
28	1	nikkei
28	2	stock
28	1	average
28	1	closed
28	2	points
28	1	up
28	1	tokyo
28	1	exchange
28	1	wednesday

TERM

term	idf
average	1.08
closed	1.08
exchange	1.00
nikkei	2.07
points	1.23
stock	1.00
tokyo	1.58
up	0.30
wednesday	0.60

LOGO

或(or)关系

Boolean Search OR query

```
    select docID
      from INDEX
    where term = term1
union
    select docID
      from INDEX
    where term = term2
 union
    select docID
      from INDEX
    where term = term3
....
union
    select docID
      from INDEX
    where term = termN
```

```
select docID
  from INDEX
where    term = term1   OR
         term = term2   OR
         term = term3   OR
         ....
         term = termN
```

LOGO

与(and)关系

Boolean Search AND query

```
    select docID
      from INDEX
    where term = term1
intersect
    select docID
      from INDEX
    where term = term2
intersect
    select docID
      from INDEX
    where term = term3
 ....
intersect
    select docID
      from INDEX
    where term = termN
```

```
select docID
  from INDEX a, INDEX b, INDEX c, ... INDEX N
where  a.term = term1        AND
       b.term = term2        AND
       c.term = term3        AND
                ....
       n.term = termN        AND
       a.docID = b.docID     AND
       b.docID = c.docID     AND
                ....
       N-1.docID = N.docID
```

Fixed Join-Count AND Queries

Find all documents that contain all of the terms found in the QUERY relation:

```
select i.docID
  from INDEX i, QUERY q
where i.term = q.term
group by i.docID
having count (distinct (i.term)) =
      select count(distinct (term)) from QUERY
```

思考题1

- Q=病毒 AND（计算机OR电脑）AND NOT 医
- 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
- 上述文档哪一个会被检索到？

思考题2

- 如果想查关于去年快乐女声5进4比赛方面的信息，用布尔模型怎么构造查询？

- (2009 OR 去年) AND (快乐女声OR 快女OR 快乐女生) AND (5进4 OR 五进四 OR 五 AND 进AND 四)
- 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。

(2009 OR 去年) AND (快乐女声OR 快女OR 快乐女生) AND (5进4 OR 五进四 OR 五AND 进AND ...

文件(F) 查看(V) 收藏(B) 帐户(U) 工具(T) 帮助(H)

后退 前进 停止 刷新 主页 恢复 收藏 历史 无痕

http://www.google.co

(2009 OR 去年)

登录 添加 谷歌 网址大全 360游戏 IE打开 充值 翻译 截图 实用工具

网页 图片 视频 地图 新闻 音乐 购物 Gmail 更多 ▼

网络历史记录 | 搜索设置 | 登录

Google

(2009 OR 去年) AND (快乐女声OR 快女OR 快乐女生) AND (5进4 搜索

获得约 15,600 条结果

高级搜索

所有结果
视频
更多

网页
所有中文网页
简体中文网页
更多搜索工具

您是不是要找： *(2009 OR 去年) AND (快乐女声OR 快女OR 快乐女声) AND (5进4 OR 五进四 OR 五AND 进AND 四)*

有关“(2009 OR 去年) AND (快乐女声OR 快女OR 快乐女生) AND (5进4 OR 五进四 OR 五AND 进AND 四)”的视频

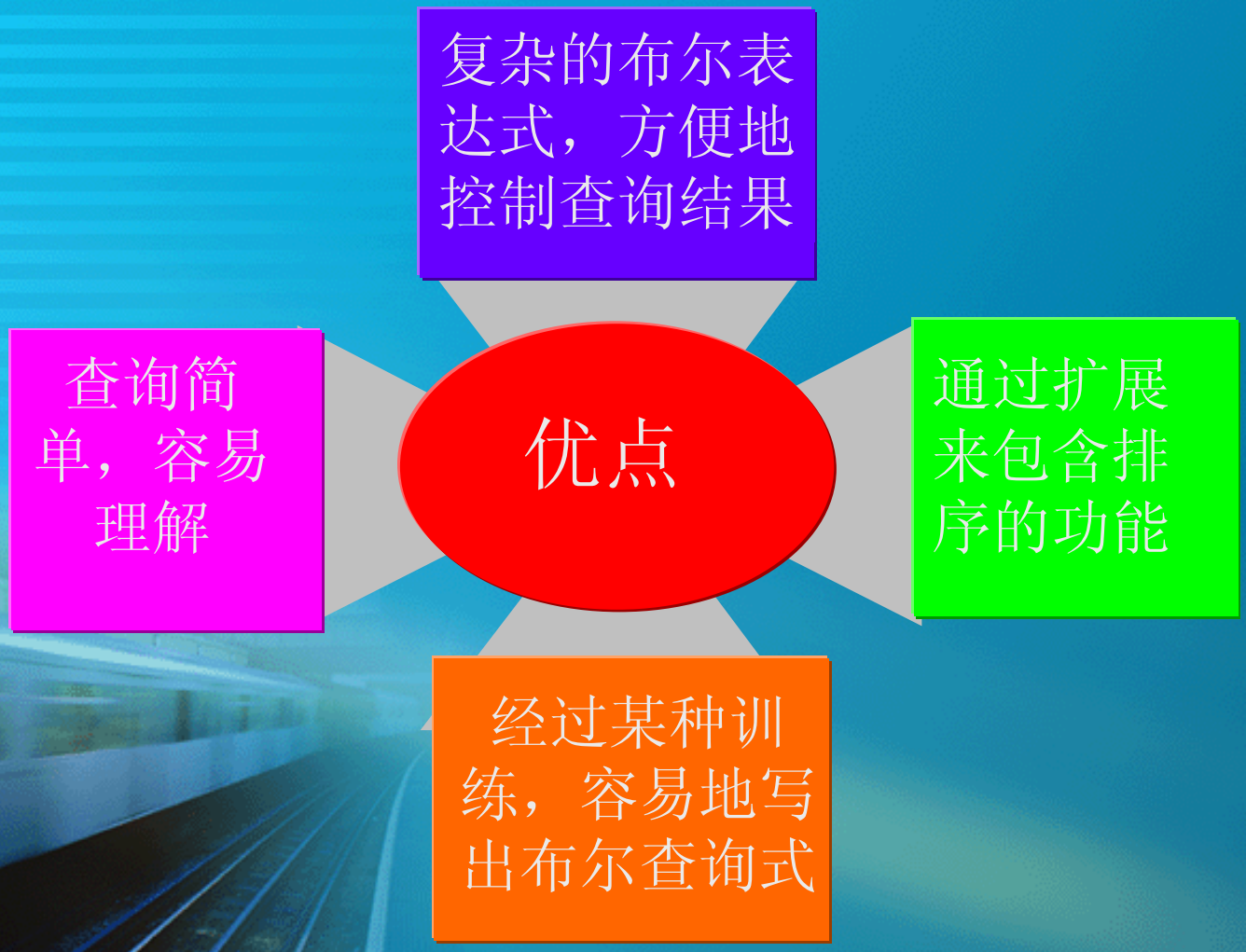
湖南卫视 2009快乐女声全国总决赛五进四淘汰赛
150 分钟 - 2009年8月21日
v.youku.com

快女 - 2009快乐女声官方赛区：湖南卫视金鹰网梦幻龙族网络直通区专区页面
湖南卫视2009年全新推出,2009快乐女声,原名,超级女声,最新的宣传片由2005年, ... 评委手记：快女总决赛5进4 非常完美的演出 • 快女五进四：留住一个她送走一个她 ...
girl09.hunantv.com/zixun/ - 网页快照 - 类似结果

2009快乐女声总决赛全部歌曲演唱会专辑mp3试听下载- 搜狗音乐大本营
2009快乐女声全部歌曲试听下载，快乐女声总决赛决战之夜、4进3、5进4、6进5， ... 来自搜狗音乐搜索中搜索量最高的2009快乐女生歌曲）. 【到“快女吧”参与网友讨论】 ...

下载 100%

3. 布尔检索优缺点


优点
复杂的布尔表达式，方便地控制查询结果
查询简单，容易理解
通过扩展来包含排序的功能
经过某种训练，容易地写出布尔查询式

只能严格匹配，很难有好的检索结果

很难控制输出量的大小

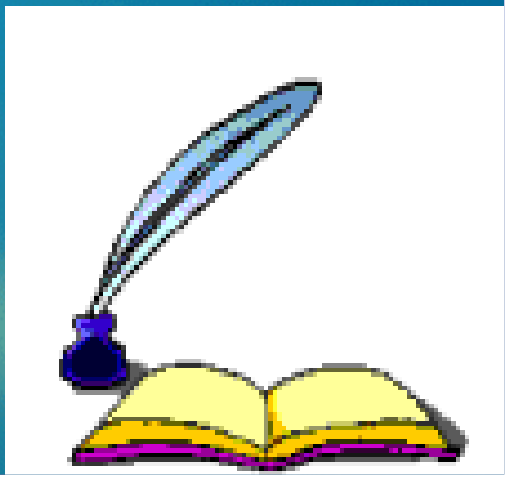
缺点

构造布尔逻辑式不容易

结果不能按用户定义的重要性排序输出

第二部分 邻近检索

邻近检索（proximity retrieval），又称为“位置检索”、“词位检索”、“全文检索”是以数据库原始记录中词语的相对次序或者位置关系为对象进行组配运算。



(W)与(nW)算符比较

W算符：
检索词必须按此前后邻接的顺序排列，顺序不可颠倒，而且检索词之间不允许有其他的词或字母，但允许有空格或连字符号

NW算符：
算符两侧的检索词之间允许插入n个实词或虚词，但两个检索词的次序还是不能颠倒。

实例

- eg:输入gas(W)condensate可检索出包含gas condensate 和gas-condensate的记录。
- eg:laser（1W）printer可检索出包含“laser printer”、“laser color printer”和“laser and printer”的记录。

(N)与(nN)算法比较

检索词彼此必须相邻接，但两个检索词的前后关系可以颠倒，两词之间不能插入任何词

- 检索词彼此必须相邻接，之间插入词的最多数目是n个，且两个检索词的次序可以任意颠倒

举例

- eg：money（N）supply可检索出包含money supply和supply money两个词组的记录；
- eg：economic(2N)recovery 可以检出包含economic recovery、recovery of the economy 、recovery from economic troubles的记录。

(S)算符

运算符两侧的检索词只要出现在记录的同一个子字段内，不限制它们在此子字段中的相对次序，中间插入词的数量也不限。

(F)算符

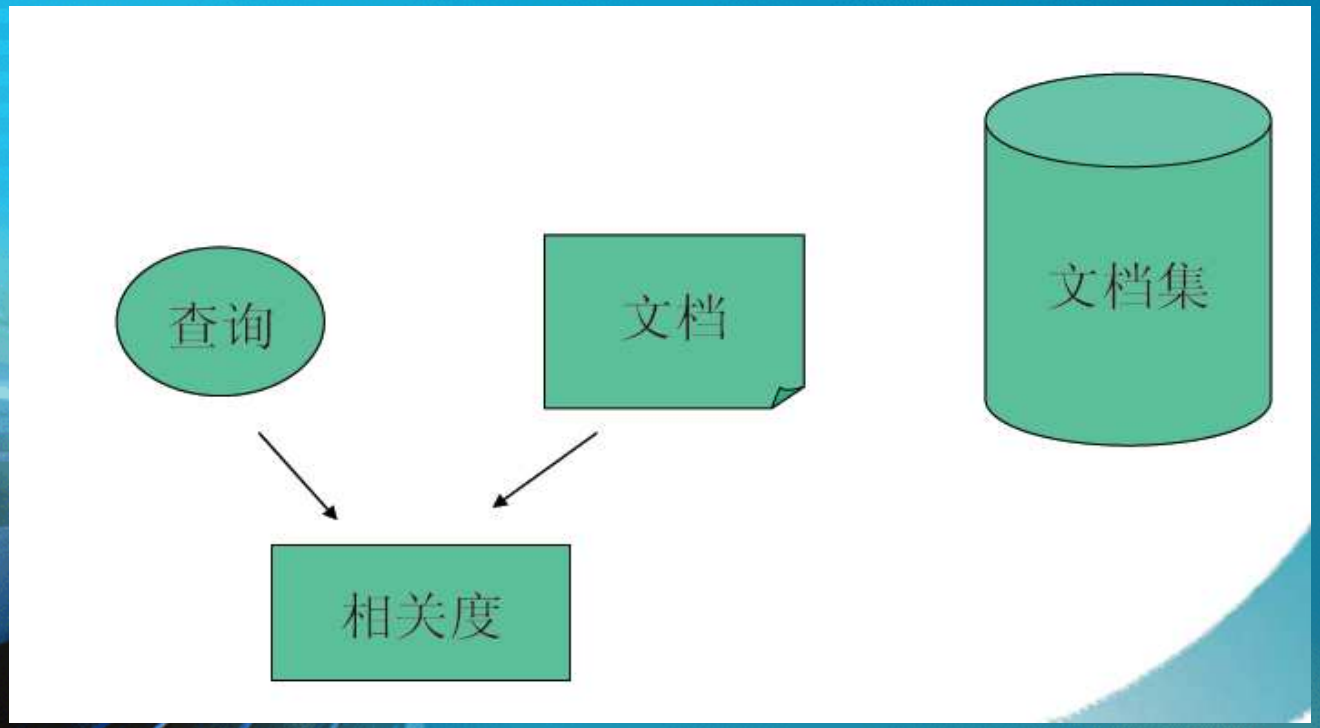
检索词必须同时出现在文献记录的统一字段内，但两个词的前后顺序不限，夹在两个词之间的词的个数也不限。

举例

- eg:“strength(S)steel”表示只要在同一句子中检索出含有“strength 和steel”形式的均为命中记录。
- eg：environmental(F) impact/DE,TI表示这两个词必须同时出现在叙词字段和篇名字段中。

第三部分 用SQL计算相关度

- 信息检索系统的目标就是检索出所有与用户查询相关的文献信息，而尽可能排除所有不相关信息。



- 形式上说，信息检索中的相关度是一个函数**R**，输入是查询**Q**、文档**D**和文档集合**C**，返回的是一个实数值。
- **R=f(Q,D,C)**
- 信息检索就是给定一个查询**Q**，从文档集合**C**中计算每篇文档**D**与**Q**的相关度并排序**(Ranking)**。

①相似度度量– 内积(Inner Product)

- 文档D 和查询式Q 可以通过内积进行计算:

$$\mathrm{sim}(D,Q) = \sum_{k=1}^{t}(d_{ik} \bullet q_k)$$

  - d_{ik} 是文档d_i中的词项k 的权重，q_k 是查询式Q中词项k的权重
- 对于二值向量，内积是查询式中的词项和文档中的词项相互匹配的数量。
- 对于加权向量，内积是查询式和文档中相互匹配的词项的权重乘积之和。

内积– 举例

- 二值（Binary）：
  - D = 1, 1, 1, 0, 1, 1, 0
  - Q = 1, 0 , 1, 0, 0, 1, 1
  - sim(D, Q) = 3

- 加权

$D_1 = 2T_1 + 3T_2 + 5T_3$, $D_2 = 3T_1 + 7T_2 + T_3$

$Q = 0T_1 + 0T_2 + 2T_3$

$sim(D_1, Q) = 2*0 + 3*0 + 5*2 = 10$

$sim(D_2, Q) = 3*0 + 7*0 + 1*2 = 2$

- 向量的大小 = 词表的大小 = 7
- 0 意味着某个词项没有在文档中出现，或者没有在查询式中出现

内积的特点

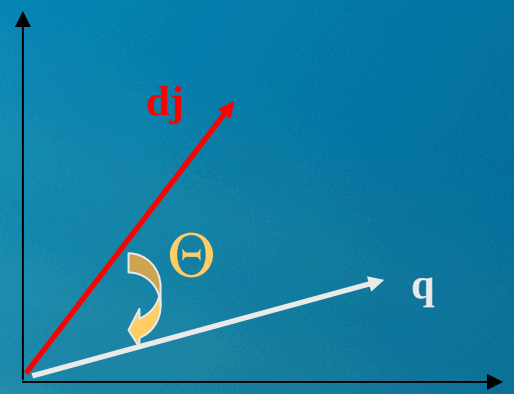
- 内积值没有界限
  - 不象概率值，要在(0,1)之间
- 对长文档有利
  - 内积用于衡量有多少词项匹配成功，而不计算有多少词项匹配失败
  - 长文档包含大量独立词项，每个词项均多次出现，因此一般而言，和查询式中的词项匹配成功的可能性就会比短文档大。

②相似度计算—夹角余弦值

相似度计算函数有很多种，较常用的是**两个向量夹角的余弦函数**。

文档和提问的相似度值由以下公式获得：

$$sim(d_j,q)=\frac{d_j \cdot q}{|d_j|*|q|}=\frac{\sum_{i=1}^{t} w_{ij}*w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2}*\sqrt{\sum_{i=1}^{t} w_{iq}^2}}$$



一个例子

- 查询*q*：(<2006,1>,<世界杯,2>)
- 文档*d1*：(<2006,1>,<世界杯,3>,<德国,1>,<举行,1>)
- 文档*d2*：(<2002,1>,<世界杯,2>,<韩国,1>,<日本，1>，<举行，1>)

	d_1	d_2	q
2002	0	1	0
2006	1	0	1
世界杯	3	2	2
德国	1	0	0
韩国	0	1	0
日本	0	1	0
举行	1	1	0

一个例子

- 查询和文档进行向量的相似度计算：
  - 采用内积

    文档d_1和q的内积：1*1+3*2=7

    文档d_2和q的内积：2*2=4
  - 夹角余弦

    文档d_1和q的夹角余弦：$\frac{7}{\sqrt{12 * 5}} = 0.9$

    文档d_2和q的夹角余弦：$\frac{4}{\sqrt{8 * 5}} \approx 0.63$

一个例子

- 查询和文档进行向量的相似度计算：
  - 采用内积

    文档d_1和q的内积：1*1+3*2=7

    文档d_2和q的内积：2*2=4
  - 夹角余弦

    文档d_1和q的夹角余弦：$\frac{7}{\sqrt{12 * 5}} = 0.9$

    文档d_2和q的夹角余弦：$\frac{4}{\sqrt{8 * 5}} \approx 0.63$

思考

- 在相似度计算的基础上，需要指定一个相似度的阀值，并规定凡与提问向量的相似度大于阀值的文档，都将作为检索结果提供给用户。这样，就可以实现“部分匹配”的思想。

Thank You !

Contact
Email: kevinzhang@bit.edu.cn

Welcome to visit my blog

http://hi.baidu.com/drkevinzhang/