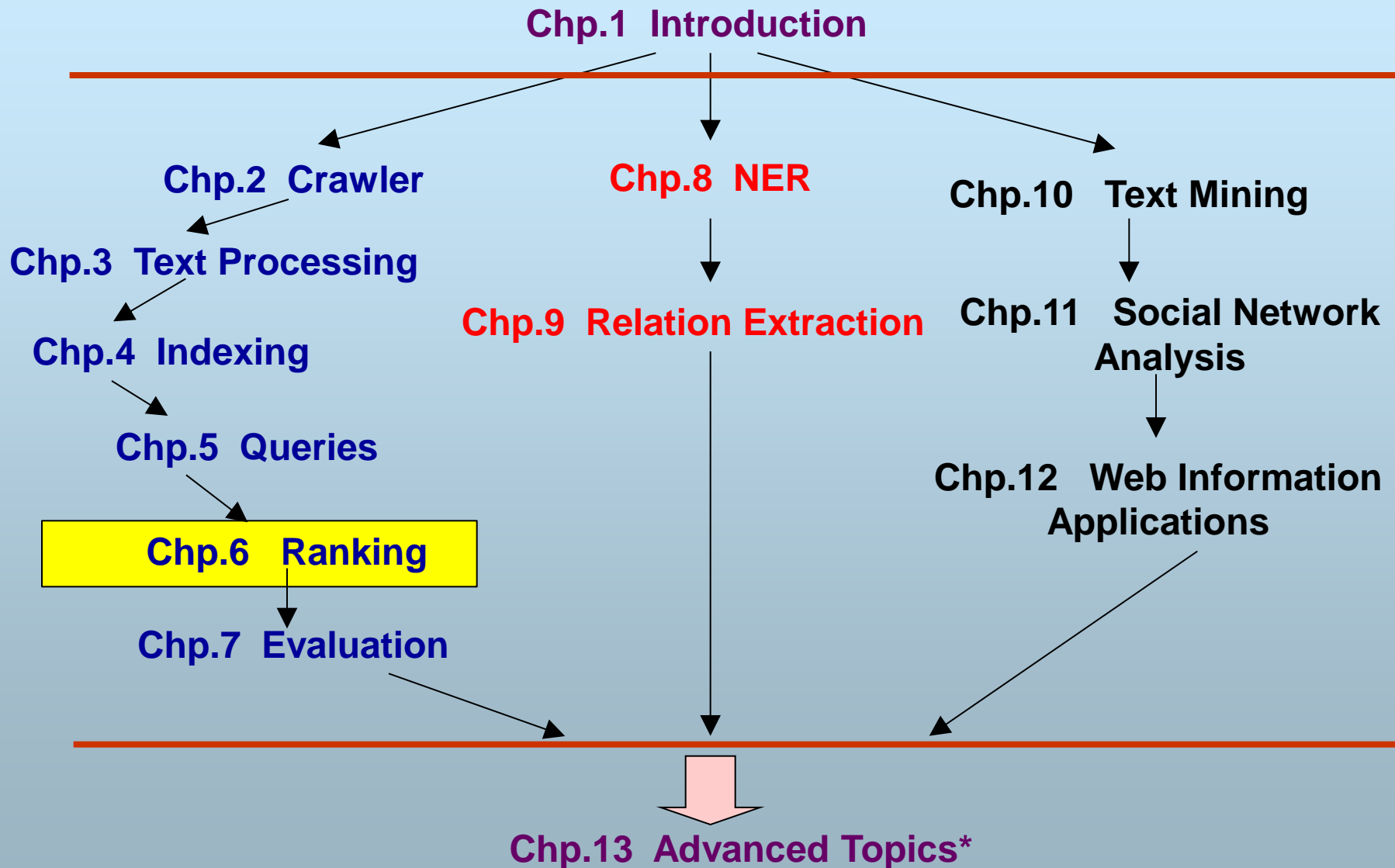


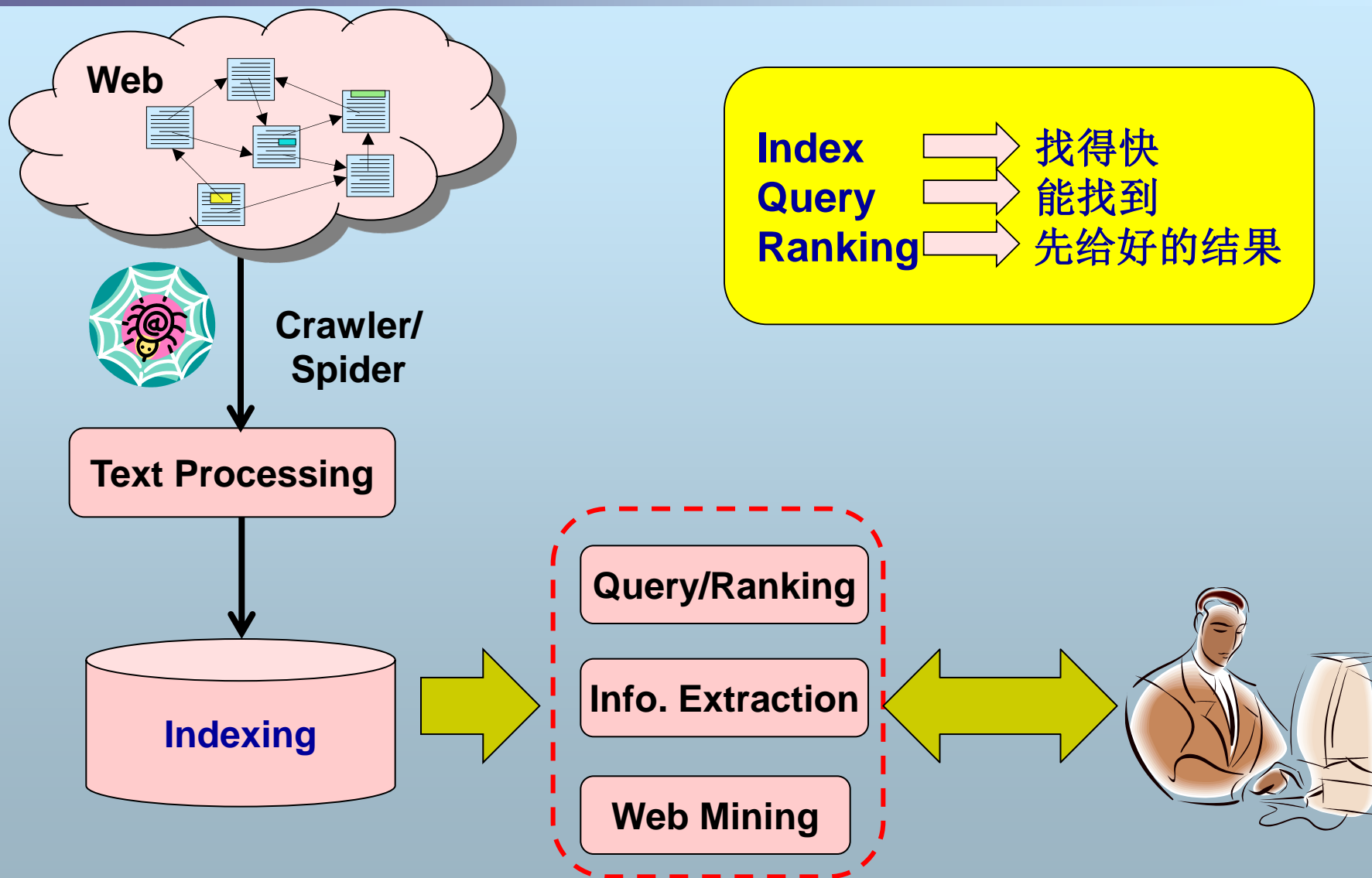
Ranking



课程知识结构



本章讨论的问题



不同的ranking算法结果

网页 找到约 199,000 条结果 (用时 0.29 秒)

中国科学技术大学
中国科学院所属的一所以前沿科学和高新技术为主、兼有以科技为背景的管理和人文 学科的综合性全国重点大学。
www.ustc.edu.cn/ - [类似结果](#)

邮箱
登录邮箱. 用户名: 密码: . 教师 学生. 界面 风格. 自动选择, AJAX 风格, HTML 风格.
email.ustc.edu.cn/

瀚海星云站
非中国教育网用户请使用本站电信线路或 网通线路. 手机用户切换到移动版 PC ...
bbs.ustc.edu.cn/

中国科学技术大学教务处
综合教务系统 - 瀚海星云站 - 考试查询 - 课表查询 - ...
www.teach.ustc.edu.cn/

University of Science and Technology of China (USTC)
The University of Science and Technology of China (USTC) is under the ...
en.ustc.edu.cn/

中国科学技术大学学位与研究生 ...
研究生信息平台 - 研究生招生在线 - 学位 - 文档下载 - ...
gradschool.ustc.edu.cn/

Baidu 百度 新闻 网页 贴吧 知道 音乐 图片 视频 地图

USTC

中国科学技术大学
少年班学院 数学科学学院 物理学院 化学与材料科学学院 生命科学学院 工程技术学院 计算机科学与技术学院 地球和空间科学学院 管理学院 公共事务
www.ustc.edu.cn/ 2013-10-12 - [百度快照](#)

ustc 百度百科

 中国科学技术大学，简称“中国科大”，隶属于中国科学院是中国科学院所属的一所以前沿科学和高新技术特色管理和人文学科的综合性全国重点大学1。1958年10月建校，1970年...
[学校简介](#) [历史沿革](#) [学校领导](#) [师资力量](#) [硬件设施](#)
baike.baidu.com/ 2013-10-13

USTC什么意思? 百度知道
1个回答 - 提问时间: 2010年09月02日
最佳答案: 这都不知道啊,世界著名大学中国科学技术大学啊!!! <http://www.uszhidao.baidu.com/link?url=vndGM70euRMwBW8...> 2010-9-2

ustc生活好多乐 是由哪些歌曲串的	1个回答	2012-02-26
ustc是什么意思?	1个回答	2013-07-18
安卓 Ustc是什么文件夹	2个回答	2012-03-27

[更多知道相关问题>>](#)

USTC生活好多乐 百度文库
★★★★★ 评分:4.5/5 2页
USTC95— A P... 0人评 4页 红歌有好多 0人评 2页 北京好多地方 0人评 2页

不同的ranking算法结果

网页 找到约 521,000 条结果 (用时 0.22 秒)

[中国科技大学](#) [Google 搜索](#) [强力驱动 Google™](#)

[欢迎蒞臨--中國科技大學--](#) [强力驱动 Google™](#)
設有規劃與設計學院、管理學院、資訊學院及人文社會學群四大學院。
www.cute.edu.tw/ - [类似结果](#)

[中國科技大學--新竹校區: 歡迎蒞臨--](#)
1 of 6. China University of Technology. 2 of 6. 賀! 本校連續第六年榮獲 教育部 補助獎勵教學卓越計畫 校、校區、區美、美景、景、青、青春、春洋、洋溢、溢、學、學習。
hcc.cute.edu.tw/ - [类似结果](#)

[中国科学技术大学.分数线.专业设置 新浪院校库 新浪教育 新浪网](#)
中国科学技术大学是中国科学院所属的一所以前沿科学和高新技术为主、兼有特色 管理和人文学科的综合性全国重点大学。1958年9月创建于北京, 首任校长由郭沫若 ...
kaoshi.edu.sina.com.cn/college/c/10358.shtml - [类似结果](#)

[中国科学技术大学](#)
中国科学院所属的一所以前沿科学和高新技术为主、兼有以科技为背景的管理和人文 学科的综合性和全国重点大学。
www.ustc.edu.cn/ - [类似结果](#)

秋季浪漫连衣裙, 让你在这个柔美的秋日里穿出时尚感, 瞬间变成柔美情人.....

[时空轨迹](#) | [个人中心](#) | [搜索设置](#) | [百度首页](#)

[新闻](#) [网页](#) [贴吧](#) [知道](#) [音乐](#) [图片](#) [视频](#) [地图](#)

[中国科技大学](#)

[中国科学技术大学](#) [官网](#)
2013年中国大学校长联谊会合肥举行C9+HK3高校共话信息化时代的大学教
美欧大学联盟签署《合肥宣言》一流大学建设研讨会大会报告精彩纷呈...
www.ustc.edu.cn/ 2013-10-12 [百度快照](#)

[中国科技大学](#) [百度百科](#)


中国科学技术大学, 简称“中国科大”, 隶属于中国科学院是中国科学院所属的一所以前沿科学和高新
特色管理和人文学科的综合性全国重点大学1。1958
1970...
[学校简介](#) [历史沿革](#) [学校领导](#) [师资力量](#) [硬件设施](#)
baike.baidu.com/ 2013-10-13

[中国科学技术大学](#) [高考招生](#) [中国教育在线](#) [高考频道](#)

学校类型: 理工类 所在地: 安徽
学校属性: [211高校](#) [985高校](#)
院校信息: [专业介绍](#) [招生章程](#) [招生计划](#)

各地区文理科历年分数线查询

考生所在地 [文科](#) [查询](#)

[理工类院校名单](#) [安徽地区院校名单](#)

Ranking的难点在哪？

■ 传统IR方法有两个重要的内在假设：

- 被索引的信息本身有很高的、同等的质量，至少在信息的组织和内容上有着较高的质量
- 检索信息的用户有一定的相关技能和知识
- E.g., 数字图书馆

■ 但这些假设在Web Search上不再成立：

- Web网页的质量参差不齐，大量的网页组织性、结构性比较差
- 大部分检索用户是没有任何经验的
- 用户的查询需求也存在着巨大差异

Ranking例子

- 如果我牙疼，应该去看怎样的医生呢？假设我有3个医生可选择：
 - **A**医生，既治眼病，又治胃病；
 - **B**医生，既治牙病，又治胃病，还治眼病；
 - **C**医生，专治牙病。
 - 假如再加一个条件：**B**医生经验丰富，有二十年从医经历，医术高明，而**C**医生只有五年从医经验。

- 结论：择医需要考虑两个条件

- 1: 医生的专长与病情的适配程度
- 2: 医生的医术

搜索引擎排序 { 网页内容与用户查询的匹配程度
网页本身的质量

本章主要内容

- IR检索模型与相关度计算
- PageRank
- HITS

一、IR检索模型与相关度计算

- 信息检索模型概述
- 布尔模型
- 向量空间模型
- 概率模型

1、信息检索模型概述

■ 信息检索模型

- 是用来描述文档和用户查询的表示形式以及它们之间相关性的框架

■ 信息检索的实质问题

- 对于所有文档，根据其与用户查询的相关程度由大到小进行排序。

■ 信息检索模型与搜索引擎排序算法关系

- 好的信息检索模型在相关性上产生和人类决策非常相关的结果，基于好的检索模型的排序算法能够在排序结果顶部返回相关的文档。

1、信息检索模型概述

■ 信息检索模型的形式化表示： $[D, Q, F, R(D_i, q)]$

- **D**: 文档表达，通常表示为索引词项的集合
- **Q**: 查询表达
- **F**: 查询与文档之间的匹配框架
- **R**: 查询与文档之间的相关性度量函数

■ 信息检索模型的分类

- 基于集合论的模型
 - ◆ 布尔模型 (**Boolean Model**)
- 基于代数论的模型
 - ◆ 向量空间模型 (**Vector Space Model**)
- 基于概率论的模型
 - ◆ 概率模型 (**Probabilistic Model**)
 - ◆ 语言模型 (**Language Model**)
 - ◆ 推理网络 (**Inference Network**)

1、信息检索模型概述


■ 相关性


- 1.相关性 {
 - 主题(**topical**)相关 (一篇文档被判定和一个查询是同一主题)
 - 用户(**user**)相关 (考虑用户在判定相关性时涉及的所有因素)
- 2.相关性 {
 - 二元(**binary**)相关 (简单判定一篇文档是相关还是非相关)
 - 多元(**multi-valued**)相关 (从多个层次判断相关性)

1、信息检索模型概述

用户相关搜索示例

DBLP FILTER Sign in

 Sort by ☐ relevance ☒ importance ☐ date

Scholar About 28 results (5.57sec)  (1998~2013)

Since Time

Since 2013

Since 2012

Since 2009

Custom range...

Sort By

Sort By Relevance

Sort By Importance

Sort By Date

A	EE	Scholar	A Data Model and Data Structures for Moving Objects Databases. (Luca Forlizzi and Ralf Hartmut G and ü) <i>ACM Conference on Management of Data (sigmod) [2000] Cited by 353</i>
A	EE	Scholar	Scientific Data Repositories: Designing for a Moving Target. (Etzard Stolte and Christoph von Praun and Gustavo Alonso) <i>ACM Conference on Management of Data (sigmod) [2003] Cited by 43</i>
A	EE	Scholar	A Data Model for Moving Objects Supporting Aggregation. (Bart Kuijpers and Alejandro A. Vaisman) <i>IEEE International Conference on Data Engineering (ICDE) [2007] Cited by 20</i>
B	EE	Scholar	Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases. (Martin Erwig and Ralf Hartmut G and ü) <i>GeoInformatica (GeoInformatica) [1999] Cited by 367</i>
B	EE	Scholar	A generic data model for moving objects. (Jianqiu Xu and Ralf Hartmut G and ü) <i>GeoInformatica (GeoInformatica) [2013]</i>
B	EE	Scholar	An Object-Field Perspective Data Model for Moving Geographic Phenomena. (Kyoung-Sook Kim and Yasushi Kiyoki) <i>Database Systems for Advanced Applications (DASFAA) [2010]</i>
C	EE	Scholar	Place: A Distributed Spatio-Temporal Data Stream Management System for Moving Objects. (Xiaopeng Xiong and Hicham G. Elmongui and Xiaoyong Chai) <i>International Conference on Mobile Data Management (MDM) [2007] Cited by 18</i>
C	EE	Scholar	An analytic solution to the alibi query in the space-time prisms model for moving object data. (Bart Kuijpers and Rafael Grimson and Walled Othman) <i>International Journal of Geographical Information Science (IJGIS) [2011] Cited by 3</i>
C	EE	Scholar	A Scaleless Data Model for Direct and Progressive Spatial Query Processing. (Sai Sun and Sham Prasher and Xiaofang Zhou) <i>International Conference on Conceptual Modeling (ER) [2004] Cited by 2</i>
C	EE	Scholar	Efficient Strip-Mode SAR Raw-Data Simulation of Fixed and Moving Targets. (Ozan Dogan and Mesut Kartal) <i>IEEE Geoscience and Remote Sensing Letters (LGRS) [2011]</i>
			Computational data modeling for network-constrained moving objects. (I aurynas Snelicvs and Christian S. Jensen and Augustas Kliovs) <i>GIS</i>

2、布尔模型

■ 最早的IR模型

- 1957年，Y.Bar-Hille就对布尔逻辑应用于计算机信息检索的可能性进行了探讨，目前仍然应用于商业系统中

■ 布尔模型的前提假设：

- 在检索到的集合中所有文档关于相关性是等价的——不考虑文档质量
- 相关性是二元的

■ 特点

- 检索的结果只输出结果（TURE | FALSE）。
- 查询项被描述为布尔逻辑操作符(AND,OR,NOT)

2、布尔模型

- 文档表示为词项集合（**Bag of Words, BOW**）
- 查询 q 被表达成索引项的布尔组合形式
 - 为方便计算文档 D 和查询 q 之间的相关度，一般将查询 q 的布尔表达式转换成析取范式（**Disjunctive Normal Form, DNF**）的形式
- **Example**
 - $q = (a \vee b) \wedge z \rightarrow (a \wedge z) \vee (b \wedge z)$

2、布尔模型

■ Example:

- $q = \text{病毒} \text{ and } (\text{计算机} \text{ or } \text{电脑}) \text{ and not 医}$
- D:
 - ◆ 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

■ 查询表示

- $q = \text{病毒} \wedge (\text{计算机} \vee \text{电脑}) \wedge \sim \text{医}$
 $= (\text{病毒} \wedge \text{计算机} \wedge \sim \text{医}) \vee (\text{病毒} \wedge \text{电脑} \wedge \sim \text{医})$

■ 采用完全匹配的方式

- If $\text{sim}(D_i, q) = 1$, 返回
- If $\text{sim}(D_i, q) = 0$, 不返回

2、布尔模型

■ 优点

- 查询项可以是任何文档的特征而非词语，易扩充，如文档日期、文档类型。
- 文档可以在搜索过程中快速被剔除。
- 结果可以预测，容易解释
- 对自身需求和文档集性质非常了解的专家而言，布尔查询是不错的选择

■ 缺点

- 效率完全依赖于用户给出的检索词
 - ◆ 大部分用户不能撰写布尔查询或者他们认为需要大量训练才能撰写合适的布尔查询
- 精确匹配，结果容易过多或过少。
- 没有相关性**ranking**处理，忽略了词项顺序、词频等因素。不太适合**Web Search**

2、布尔模型

- 布尔查询常常会得到过少($=0$)或者过多(>1000)的结果
 - 查询 1 (布尔与操作): [iPhone IOS]
 - ◆ \rightarrow 200,000 个结果 – 太多
 - 查询2 (布尔与操作): [iPhone IOS newest]
 - ◆ \rightarrow 0 个结果 – 太少
- 在布尔检索中，需要大量技巧来生成一个可以获得合适规模结果的查询

3、向量空间模型

■ tf-idf

■ 向量空间模型

- 每篇文档表示成一个基于**tf-idf**权重的实值向量 $\mathbf{D} \in \mathbb{R}^{|V|}$.
- 查询也同样表示为词项的**tf-idf**权重向量
- 通过文档-查询的相似度（邻近度）返回结果并排序

$$\begin{aligned} d_j &= (w_{1,j}, w_{2,j}, \dots, w_{t,j}) \\ q &= (w_{1,q}, w_{2,q}, \dots, w_{t,q}) \end{aligned}$$

Donna Harman: **Relevance Feedback Revisited**. SIGIR 1992: 1-10

Gerard Salton, A. Wong, C. S. Yang: **A Vector Space Model for Automatic Indexing**. Commun.

ACM 18(11): 613-620 (1975)

相关性排序

■ 我们希望

- 在同一查询下，文档集中相关度高的文档排名高于相关度低的文档

■ 如何实现？

- 通常做法是对每个查询-文档对赋一个 $[0, 1]$ 之间的分值
- 该分值度量了文档和查询的匹配程度

第一种方法：Jaccard系数

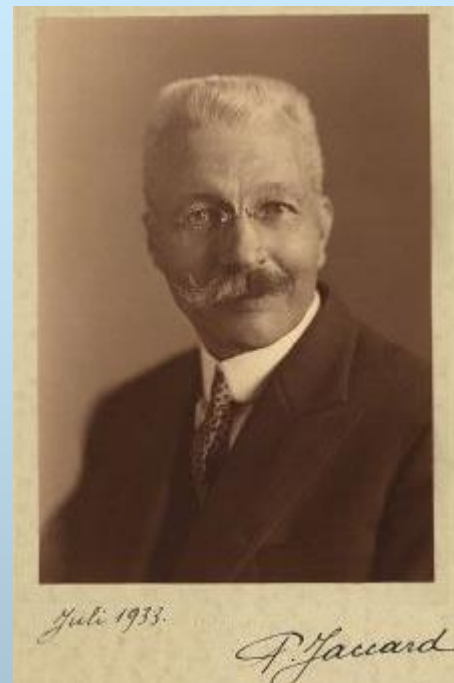
■ 1901年Jaccard提出的计算两个集合重合度的常用方法

- 令 **A** 和 **B** 为两个集合
- Jaccard系数的计算方法:

$$\text{JACCARD}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- $\text{JACCARD}(A, A) = 1$
- $\text{JACCARD}(A, B) = 0$ 如果 $A \cap B = 0$

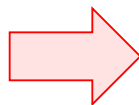
■ Jaccard 系数会给出一个0到1之间的值



Paul Jaccard (1868-1944)

查询 “ides of March”

文档 “Caesar died in March”



$$\text{JACCARD}(q, d) = 1/6$$

第一种方法：Jaccard系数

■ Jaccard系数的缺点

- 不考虑词项频率，即词项在文档中的出现次数
- 没有仔细考虑文档的长度因素
- 罕见词比高频词的信息量更大，Jaccard系数没有考虑这个信息

香农（C. E. Shannon）信息论

信息量是指从N个相等可能事件中选出一个事件所需要的信息度量或含量，也就是在辩识N个事件中特定的一个事件的过程中所需要提问“是或否”的最少次数。香农信息论应用概率来描述不确定性。信息是用不确定性的量度定义的。

一个消息的可能性愈小，其信息量愈多；而消息的可能性愈大，则其信息愈少。事件出现的概率小，不确定性越多，信息量就大，反之则少。

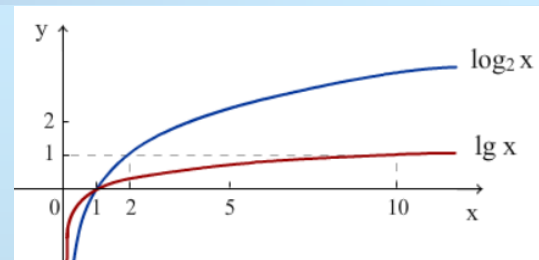
第二种方法：词项频率tf

- 词项 t 的词项频率 $tf(t,d)$ 是指词项 t 在文档 d 中出现的次数
- 利用 tf 来计算文档评分的方法
 - 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100次，即 $tf = 100$ ，在B文档中 $tf = 10$ ，那么A比B更相关
 - ◆ 但是相关度不会相差10倍，相关度不会正比于词项频率 tf

第二种方法：词项频率tf

- 另一种方法：使用对数词频
- **t** 在 **d** 中的对数词频权重定义如下：

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



- $\text{tf}_{t,d} \rightarrow w_{t,d}$:
 - $0 \rightarrow 0, 1 \rightarrow 1, 2 \rightarrow 1.3, 10 \rightarrow 2, 1000 \rightarrow 4$, 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 $\sum_{t \in q \cap d} (1 + \log \text{tf}_{t,d})$
- 如果两者没有公共词项，则得分为0

第三种方法：文档频率df

- 词项的文档频率：出现词项的文档数目
- 按照香农信息论，罕见词项比常见词所蕴含的信息更多
 - 考虑查询中某个词项，它在整个文档集中非常罕见（例如 **arachnoid**），某篇包含该词项的文档很可能相关
 - 因此，我们希望罕见词项有较高权重
 - 常见词项频繁出现在文档集中（如 **dog, cat, food** 等等），这些词对于相关度而言并不是非常强的指示词，因此频繁词会给一个正的权重，但是这个权重小于罕见词权重

第三种方法：文档频率df

- 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(Inverse document frequency):

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

(其中N 是文档集中文档的数目)

- idf由剑桥大学的Spock Jones于1972年提出，Salton在之后对tf-idf做了大量推广工作
- idf_t 是反映词项t的信息量的一个指标
- 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

第三种方法：文档频率df

■ idf计算示例

$$\text{idf}_t = \log_{10} \frac{1,000,000}{\text{df}_t}$$

词项	df_t	idf_t
calpurnia	1	6
animal	100	4
sunday	1000	3
fly	10,000	2
under	100,000	1
the	1,000,000	0

第四种方法：tf-idf

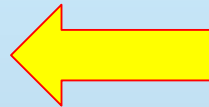
- 词项的tf-idf权重是tf权重和idf权重的乘积

$$w_{t,d} = (1 + \log \text{tf}_{t,d}) \cdot \log \frac{N}{\text{df}_t}$$

- 随着词项频率的增大而增大
- 随着词项罕见度的增加而增大
- 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 *idf、tfidf等

3、向量空间模型

- tf-idf
- 向量空间模型



3、向量空间模型

- 向量空间模型（Vector Space Model, VSM）是由G.Salton等人在1958年提出的。代表系统SMART
 - $D=\{D_1, D_2, \dots\}$, $D_i=(W_{i1}, W_{i2}, \dots, W_{in})$, W_{ij} 是词项的tf-idf权值
 - $q=(W_{q1}, W_{q2}, \dots, W_{qn})$, W_{qi} 是查询词项的tf-idf值
 - F: 非完全匹配方式
 - R
 - ◆ 用文档和查询两个向量相似度来估计文档和查询的相关性
 - ◆ 文档和查询之间的相关度具有较强的可计算性和可操作性, 不再只有0和1两个值
- 前提假设
 - 在检索到的集合中所有文档关于相关性是不等价的。
 - 相关性是多元的。
 - 查询关键字之间是相互独立的。

3、向量空间模型

■ 文档表示为向量

- 每篇文档表示成一个基于**tf-idf**权重的实值向量 $\mathbf{d} \in \mathbb{R}^{|V|}$.
 - ◆ $|V|$ 维实值空间
- 空间的每一维都对应词项
- 文档都是该空间下的一个点或者向量
- 对于**Web**搜索引擎，空间维数会非常大
- 对每个向量来说又非常稀疏，大部分都是0

3、向量空间模型

■ 文档表示示例

	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
WORSE	1.37	0.0	0.11	4.15	0.25	1.95
...						

$D1=(5.25, 1.21, 8.59, 0.0, 2.85, 1.51, 1.37)$

3、向量空间模型

■ 查询的表示

- 对于查询做同样的处理，即将查询表示成同一高维空间的tf-idf向量
- 按照文档对查询的邻近程度排序
 - ◆ 通过邻近度来度量相似度
 - ◆ 邻近度 \approx 距离的反面

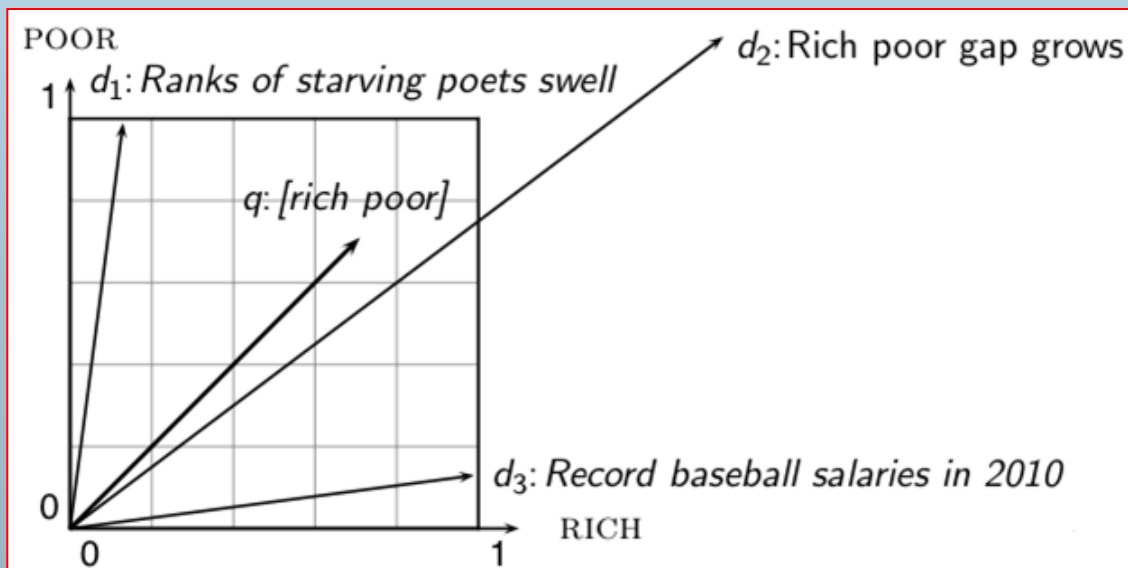
3、向量空间模型

■ 查询与文档的相似度计算

- 一种方法是采用欧氏距离，相似度为距离的倒数

$$\text{dist}(\vec{x}, \vec{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}$$

- 但是，欧氏距离不是一种好的选择，这是因为欧氏距离对向量长度很敏感

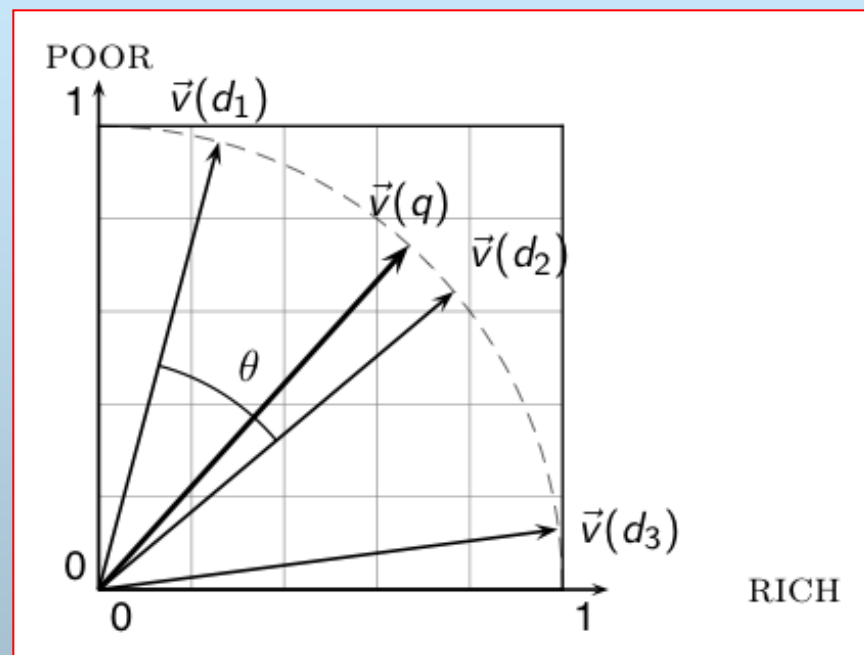


尽管查询 q 和文档 d_2 的词项分布非常相似，但是采用欧氏距离计算它们对应向量之间的距离非常大

3、向量空间模型

■ 余弦相似度

- 按文档向量和查询向量的夹角大小来计算相似度
- 假想实验
 - ◆ 将文档 d 复制一份加在自身末尾得到文档 d' ， d' 是 d 的两倍
 - ◆ 很显然，从语义上看， d 和 d' 具有相同的内容
 - ◆ 两者之间的夹角为 0 ，代表它们之间具有最大的相似度
 - ◆ 但是，它们的欧氏距离可能会很大



3、向量空间模型

■ 余弦相似度

- 向量夹角转换为夹角余弦

- ◆ 按照夹角从小到大排列文档
= 按照余弦从大到小排列文档

- 因为在区间 $[0^\circ, 180^\circ]$ 上，余弦函数**cosine**是一个单调递减函数

- 余弦定理计算**cosine**

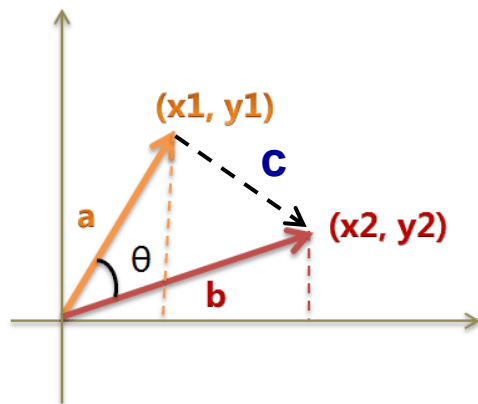
$$\cos \alpha = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| \times |\vec{y}|}$$

$$|\vec{x}| = \|\vec{x}\| = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$$

$$\cos \theta = \frac{a^2 + b^2 - c^2}{2ab}$$



$$\cos \theta = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}}$$



3、向量空间模型

■ 余弦相似度计算

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

- 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

3、向量空间模型

■ 余弦相似度计算举例

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

● 2个文档 D_1, D_2 和1个查询 Q

◆ $D_1 = (0.5, 0.8, 0.3), D_2 = (0.9, 0.4, 0.2), Q = (1.5, 1.0, 0)$

$$\begin{aligned} \text{Cosine}(D_1, Q) &= \frac{(0.5 \times 1.5) + (0.8 \times 1.0)}{\sqrt{(0.5^2 + 0.8^2 + 0.3^2)(1.5^2 + 1.0^2)}} \\ &= \frac{1.55}{\sqrt{(0.98 \times 3.25)}} = 0.87 \\ \\ \text{Cosine}(D_2, Q) &= \frac{(0.9 \times 1.5) + (0.4 \times 1.0)}{\sqrt{(0.9^2 + 0.4^2 + 0.2^2)(1.5^2 + 1.0^2)}} \\ &= \frac{1.75}{\sqrt{(1.01 \times 3.25)}} = 0.97 \end{aligned}$$

3、向量空间模型

■ 向量空间模型总结

- 将每篇文档表示成词项的**tf-idf**权重向量
- 将查询表示成词项的**tf-idf**权重向量
- 计算两个向量之间的某种相似度(如余弦相似度)
- 按照相似度大小将文档排序
- 将**Top-K**（如**K=10**）篇文档返回给用户

3、向量空间模型

■ 优点:

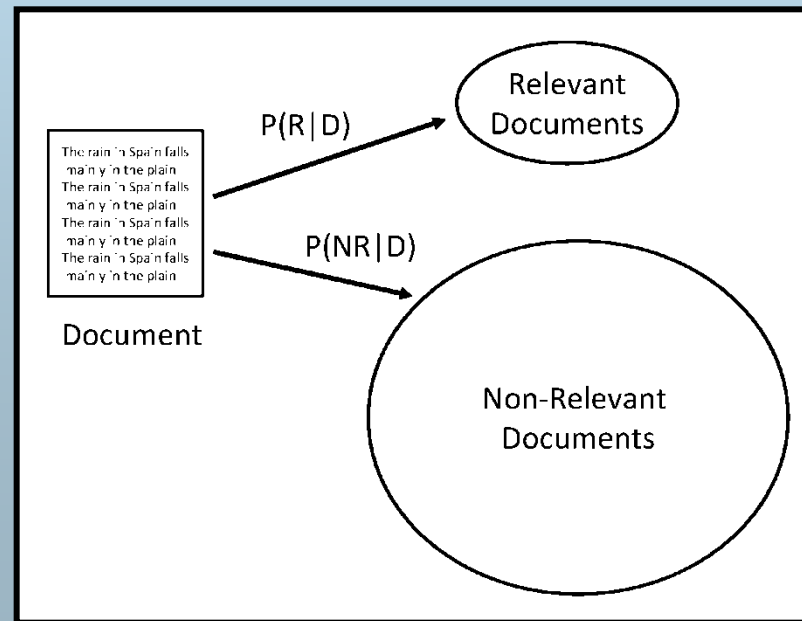
- 简洁直观，可以应用到很多其他领域(如文本分类)
- 可以支持不同的相似性度量方法以及词项权值计算方法
- 实际应用中检索效果不错

■ 缺点:

- 用户无法描述词项之间的关系
 - ◆ 例如“**world cup**”查询会返回只涉及**world**或**cup**的结果
- **tf-idf**值高的词项可能会在检索中影响过大
- 词项之间的独立性假设与实际不符
 - ◆ 实际上，词项的出现之间是有关系的，不是完全独立的。如：“王励勤”和“乒乓球”的出现不是独立的。

4、概率模型

- 按照文档与给定查询的相关性概率大小进行文档排序
 - $P(\text{relevant} \mid \text{document}_i, \text{query})$
- 通过概率的方法将查询和文档联系起来
 - 定义3个随机变量R、Q、D：相关度 $R=\{0,1\}$ ，查询 $Q=\{q_1, q_2, \dots\}$ ，文档 $D=\{d_1, d_2, \dots\}$ ，则可以通过计算条件概率 $P(R=1 \mid Q=q, D=d)$ 来度量文档和查询的相关度。



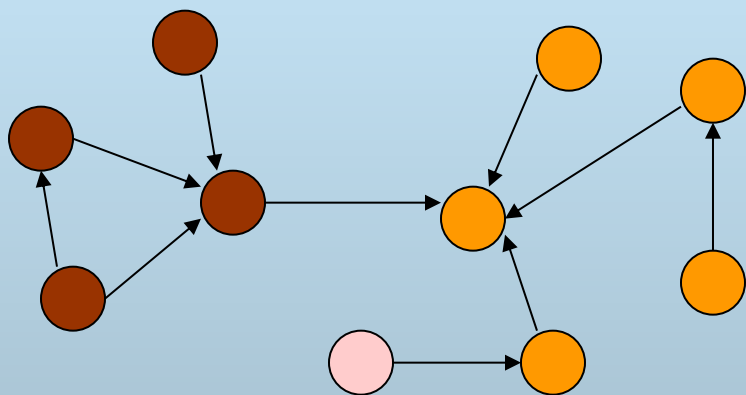
本章主要内容

- IR检索模型与相关度计算
- PageRank 
- HITS

二、PageRank

■ PageRank算法

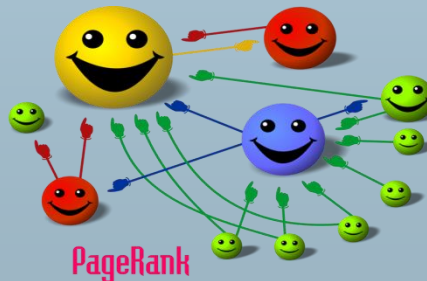
- 将网页或者文档视作一个点，网页之间的超链接视作有向边，则会构成一个巨大的有向图。



- 相同颜色代表主题相关网页（主题相关的点的连接要多于普通网页之间的连接），点之间的有向连接反映了网页之间互相引用，参考和推荐的关系，入度越多，则被引用或推荐的次数越多，网页的重要性就越大

1、PageRank历史

- Sergey Brin和Lawrence Page在1998年提出了PageRank算法，同年Jon Kleinberg提出了HITS算法
 - Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, 'The PageRank Citation Ranking: Bringing Order to the Web', 1998. [citation: 10k+]
<http://www-db.stanford.edu/~backrub/pageranksub.ps>
 - Sergey Brin, Lawrence Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine. WWW'98, Computer Networks 30(1-7): 107-117 (1998)
- PageRank(TM) 是美国 Google 公司的登记注册商标。



2、PageRank计算公式

■ PageRank的核心公式为：

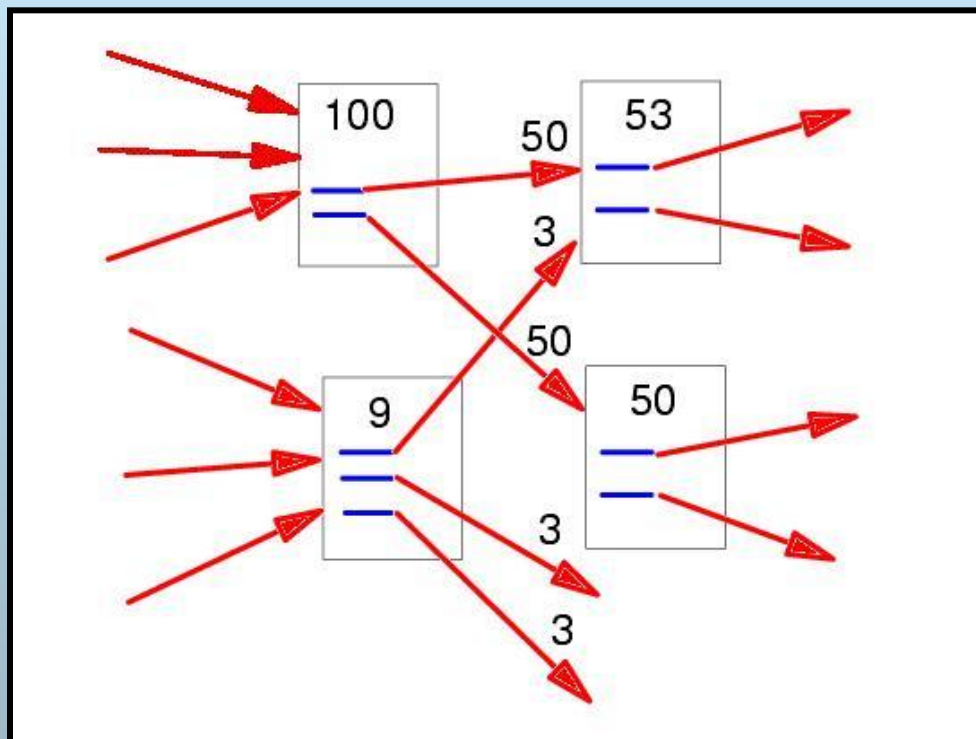
$$PR(a) = (1 - d) + d(PR(T_1) / C(T_1) + \dots PR(T_n) / C(T_n))$$

■ 其中

- **PR(a)**: 页面a的PageRank值;
- **PR(T1)**: 页面T1的PageRank值, 页面T1链向a;
- **C(T1)**: 页面T1链出的链接数量;
- **d**: 阻尼系数, 取值在0-1之间

3、PageRank背后的idea

- PageRank 是基于「从许多优质的网页链接过来的网页，必定还是优质网页」的回归关系，来判定所有网页的重要性。



- 反向链接数 (单纯的意义上的受欢迎度指标)

- 反向链接是否来自推荐度高的页面 (有根据的受欢迎指标)

- 反向链接源页面的链接数 (被选中的几率指标)

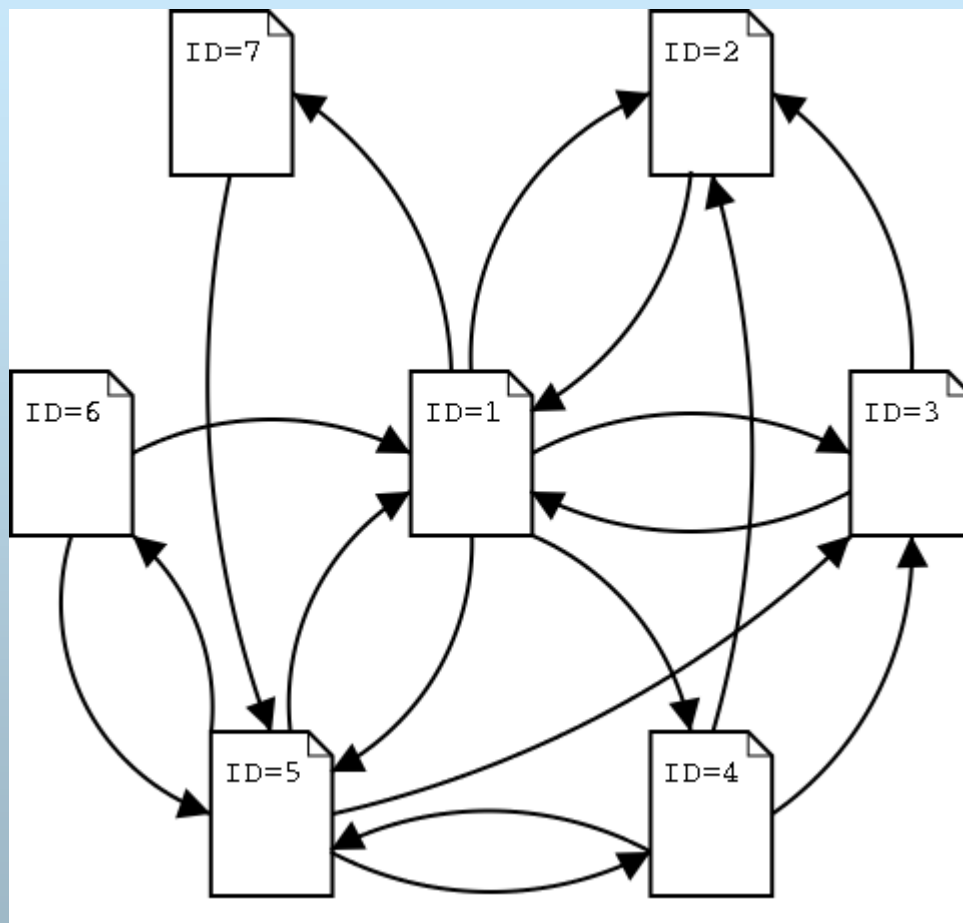
因此，如果被Yahoo! 这种PageRank 非常高的站点链接的话，网页的PageRank 也会一下子上升；相反地，无论有多少反向链接数，如果全都是从那些没有多大意义的页面链接过来的话，PageRank 也不会轻易上升。

4、PageRank的迭代计算过程

- **Google**采用了一种近似的迭代方法计算网页级别，即先给每个网页赋予一个初值，然后利用上面的公式，循环进行有限次运算得到近似的网页级别
- **Sergey Brin and Lawrence Page**的论文显示，实际进行大约**100**次迭代才能得到整个网络的网页级别。中等规模的网站计算**26,000,000**网页的**PageRank**值要花费几小时

5、PageRank计算示例

链接源ID	链接目标ID
1	2,3,4,5,7
2	1
3	1,2
4	2,3,5
5	1,3,4,6
6	1,5
7	5



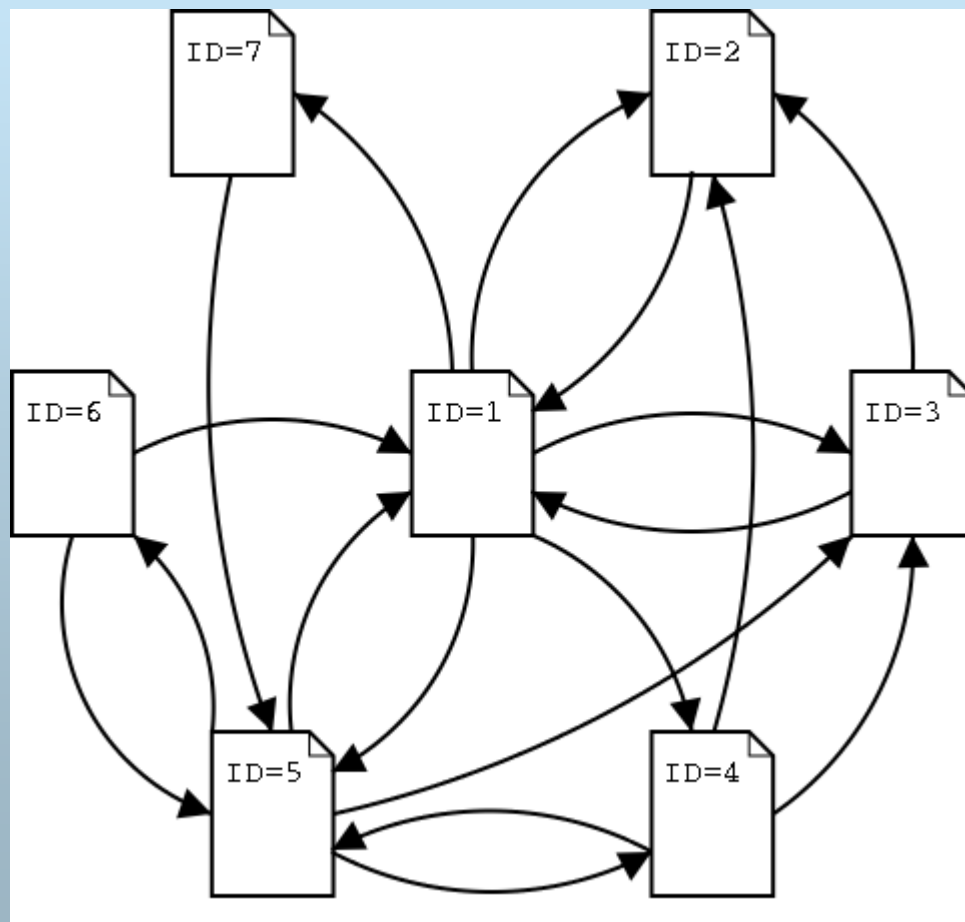
5、PageRank计算示例

■ 构造链接矩阵A

$A_{ij}=1$ if (页面 i 链接页面 j) otherwise 0

A=

```
0, 1, 1, 1, 1, 0, 1;  
1, 0, 0, 0, 0, 0, 0;  
1, 1, 0, 0, 0, 0, 0;  
0, 1, 1, 0, 1, 0, 0;  
1, 0, 1, 1, 0, 1, 0;  
1, 0, 0, 0, 1, 0, 0;  
0, 0, 0, 0, 1, 0, 0;
```



5、PageRank计算示例

- R: 将 A 倒置后将各个数值除以各自的非零元素数

$$R = \begin{bmatrix} 0, & 1, & 1/2, & 0, & 1/4, & 1/2, & 0; \\ 1/5, & 0, & 1/2, & 1/3, & 0, & 0, & 0; \\ 1/5, & 0, & 0, & 1/3, & 1/4, & 0, & 0; \\ 1/5, & 0, & 0, & 0, & 1/4, & 0, & 0; \\ 1/5, & 0, & 0, & 1/3, & 0, & 1/2, & 1; \\ 0, & 0, & 0, & 0, & 1/4, & 0, & 0; \\ 1/5, & 0, & 0, & 0, & 0, & 0, & 0; \end{bmatrix}$$

5、PageRank计算示例

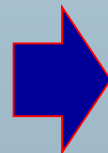
- 现有7个文档，假设初始的rank值均为1/7

$$R = \begin{bmatrix} 0, & 1, & 1/2, & 0, & 1/4, & 1/2, & 0; \\ 1/5, & 0, & 1/2, & 1/3, & 0, & 0, & 0; \\ 1/5, & 0, & 0, & 1/3, & 1/4, & 0, & 0; \\ 1/5, & 0, & 0, & 0, & 1/4, & 0, & 0; \\ 1/5, & 0, & 0, & 1/3, & 0, & 1/2, & 1; \\ 0, & 0, & 0, & 0, & 1/4, & 0, & 0; \\ 1/5, & 0, & 0, & 0, & 0, & 0, & 0; \end{bmatrix}$$

$$M = \begin{bmatrix} 1/7 \\ 1/7 \\ 1/7 \\ 1/7 \\ 1/7 \\ 1/7 \\ 1/7 \end{bmatrix}$$

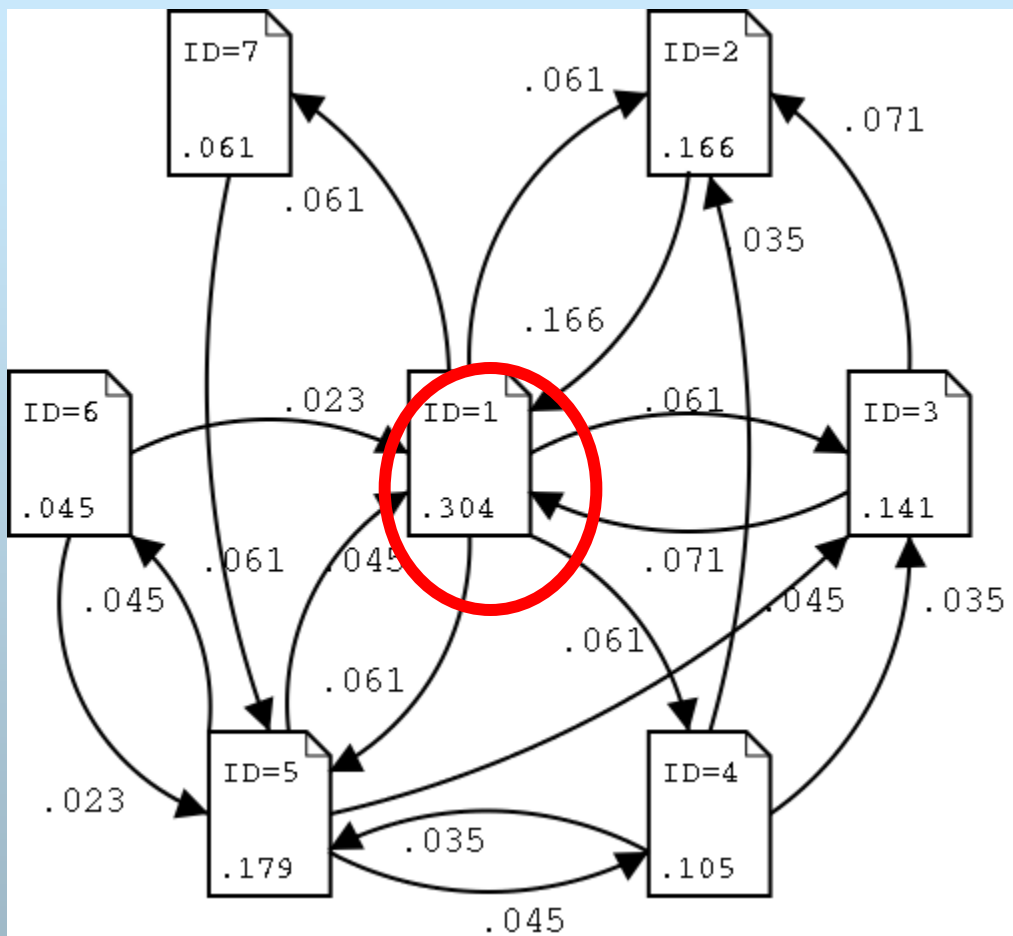
- R的第1行即为文档D1，D2，...，D7链接到D1的概率，R第1行乘以M第一列，即为D1新的rank值，同理可计算其它文档的值

$$R \cdot M = \begin{bmatrix} 0.321 \\ 0.148 \\ 0.148 \\ 0.064 \\ 0.148 \\ 0.036 \\ 0.029 \end{bmatrix} = M'$$



M'和M相差高于阈值，因此继续迭代计算 $R \cdot M'$ 得到新的M'。直到新的rank值差别小于某个阈值。

5、PageRank计算示例



- ID=1的文档流入量
 =(ID=2发出的Rank)
 +(ID=3发出的Rank) PR值/出链数
 +(ID=5发出的Rank)
 +(ID=6发出的Rank)
 = 0.166+0.141/2+0.179/4+0.045/2
 = 0.30375

递归计算

$$\forall_v Rank_{i+1}(v) = \sum_{u \in B_v} Rank_i(u) / N_u$$

递归结束标志: $|Rank_{i+1} - Rank_i| < \text{阈值}$

Larry Page和**Sergey Brin** 两人从理论上证明了不论初始rank值如何选取, 这种算法都保证了网页排名的估计值能收敛到他们的真实值。

6、PageRank的总结

■ 基于链接分析的全局网页排序算法

■ 优点

- 对互联网上的网页给出了一个全局的重要性排序，并且算法的计算过程是可以离线完成的，这样有利于迅速响应用户的请求
- 独立于检索主题，因此也常被称为**Query-independent**算法

■ 缺点

- 主题无关性，没有区分页面内的导航链接、广告链接和功能链接等，容易对广告页面有过高评价
- 旧网页容易比新网页排名高，因为新网页通常不会有很多入链
- 一般不能单独用于排序，需要跟内容排序方法相结合（例如：**tf-idf**）

三、HITS

■ **PageRank**算法中对于向外链接的权值贡献是平均的，也就是不考虑不同链接的重要性。而**Web**的链接具有以下特征：

- 有些链接具有注释性，也有些链接是起导航或广告作用。有注释性的链接更具有权威性。
- 基于商业或竞争因素考虑，很少有**Web**网页指向其竞争领域的权威网页。
- 权威网页很少具有显式的描述，比如**Google**主页不会明确给出**Web**搜索引擎之类的描述信息。

■ 总之，平均的分布权值不符合链接的实际情况

1、HITS背景

- 康奈尔大学(**Cornell University**) 的Jon Kleinberg 博士于1998 年首先提出。Kleinberg 认为既然搜索是开始于用户的检索提问，那么每个页面的重要性也就依赖于用户的检索主题。
- **HITS** (**Hyperlink - Induced Topic Search**)

1. Jon M. Kleinberg: **Authoritative Sources in a Hyperlinked Environment**. [SODA 1998](#): 668-677 (1998)
2. Jon M. Kleinberg: **Authoritative Sources in a Hyperlinked Environment**. [J. ACM 46\(5\)](#): 604-632 (1999) (Journal version)

2、权威网页与Hub网页

- **权威（Authority）网页**：与某个领域或者某个话题相关的高质量网页。
 - 比如搜索引擎领域，**Google**和**百度**首页即该领域的高质量网页，比如视频领域，**优酷**和**土豆**首页即该领域的高质量网页
- **Hub网页**：包含了很多指向高质量“**Authority**”页面链接的网页
 - **hao123**首页可以认为是一个典型的高质量“**Hub**”网页。
- **HITS**算法的目的即是在海量网页中找到与用户查询主题相关的高质量“**Authority**”页面和“**Hub**”页面，尤其是“**Authority**”页面，因为这些页面代表了能够满足用户查询的高质量内容

3、HITS算法假设

■ 基本假设1

- 一个好的“**Authority**”页面会被很多好的“**Hub**”页面指向；

■ 基本假设2:

- 一个好的“**Hub**”页面会指向很多好的“**Authority**”页面；

■ 在HITS算法中,对每个网页都要计算两个值: 权威值(authority)与中心值(hub)

3、Hub和Authority计算

计算过程有两个步骤：**I**步骤和**O**步骤。在**I**步骤中每个页面的**Authority**值为所有指向它的页面的**Hub**值之和。在**O**步骤中每个页面的**Hub**值为所有它指向的页面的**Authority**值之和。定义如下：

I 步骤:

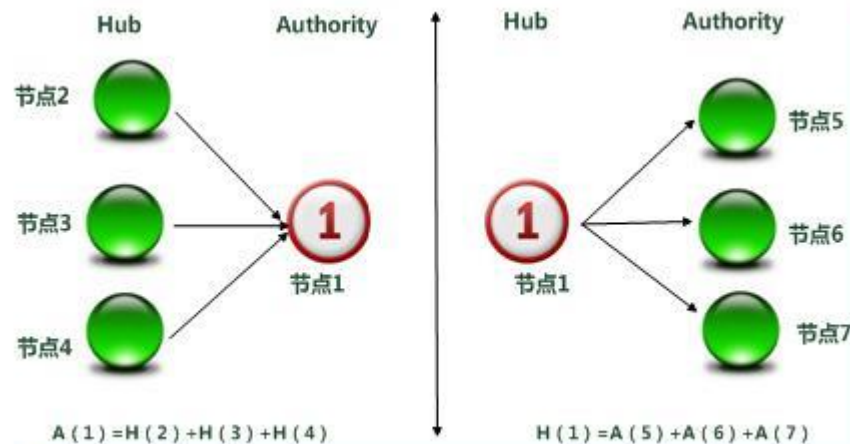
$$a_i = \sum_{j \in B(i)} h_j$$

O 步骤:
$$h_i = \sum_{j \in F(i)} a_j$$

• 初始时: $h(v)=a(u)=1$

HITS算法迭代计算这两个步骤，直到两个数值最后收敛为止。

HITS算法是在线计算的，算法的开销比较大，对每次的查询都需要临时的计算。



4、HITS算法评价

- 若一个网页有很多好的**Hub**指向，则其权威值会相应增加(即权威值增加为所有指向它的网页的现有**Hub**值之和)
- 若一个网页指向许多好的权威页，则**Hub**值也会相应增加(即**Hub**值增加为该网页链接的所有网页的权威值之和)
- **HITS**算法输出一组具有较大**Hub**值的网页和具有较大权威值的网页。

5、HITS算法总结

■ 基于权威度和中心度的网页排序算法

■ 优点

- 能更好地描述互联网的组织特点
- 主题相关, **Query-dependent**
- 可以单独用于网页排序

■ 缺点

- 需要在线计算, 时间代价较大
- 容易受到“链接作弊”的影响
 - ◆ 例如, 你可以做一个网页链接许多领域内所有的权威网页, 就成为了一个**hub**网页, 然后在该页面里放个链接指向自己的一个普通网页, 此网页**authority**值会大大提升
- 稳定性较差, 容易被链接结构变化影响



本章小结

- IR检索模型与相关度计算
 - tf-idf计算、余弦相似度
- PageRank
- HITS

实际系统中，往往需要结合多种**Ranking**的算法，构建基于组合特征的排序算法