

信息检索与Web搜索

第13讲 Web搜索基础

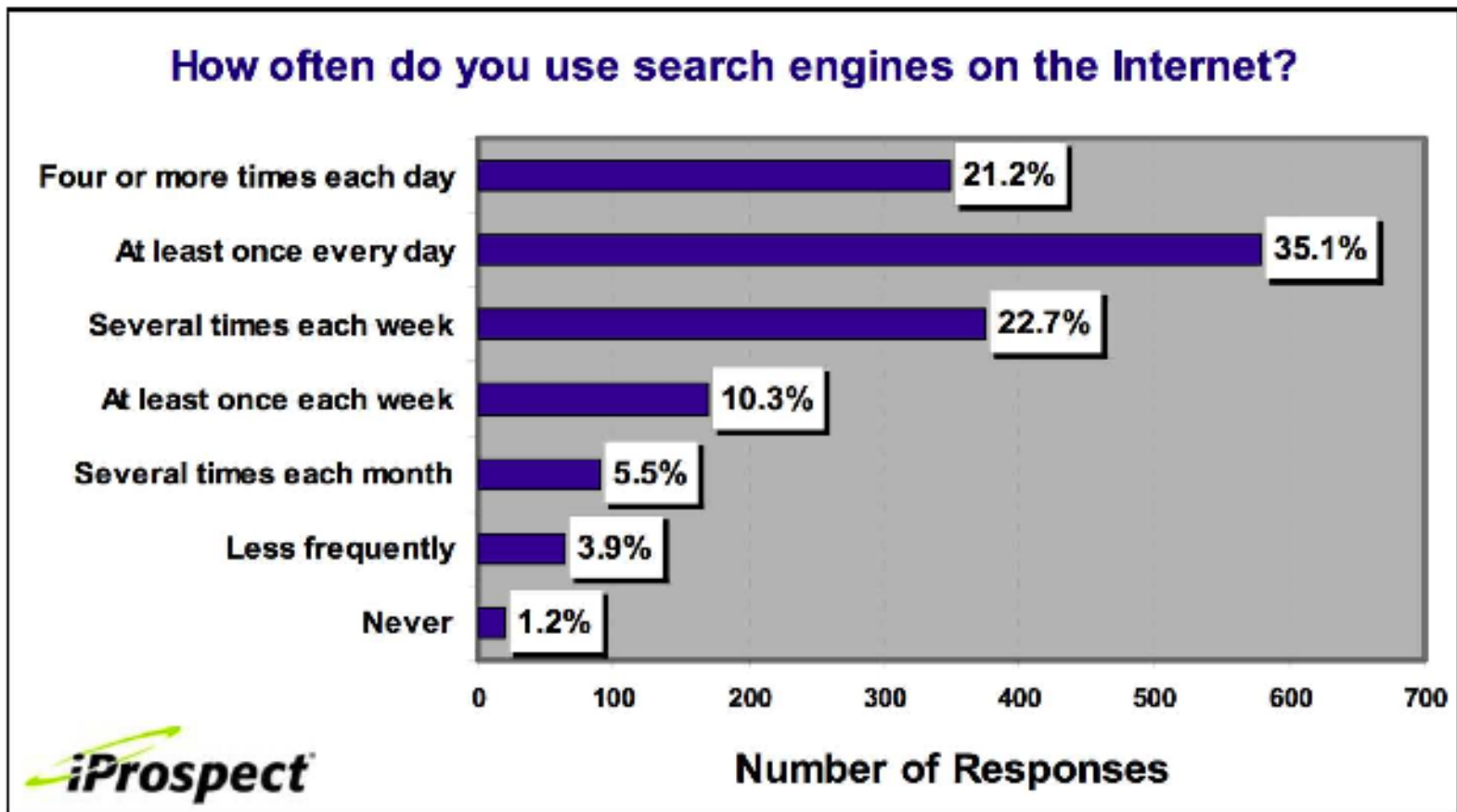
Web Search Basics

授课人：高曙明

背景与历史

- Web是提供和消费各种信息的重要场所
- Web信息发现的两种方式
 - Web网页分类体系 (Yahoo!)
 - 基于全文索引的搜索引擎 (Altavista)
- Web网页分类体系存在的问题
 - 需要人工编辑，难以扩展
 - 体系越来越大，用户体验不易保证
- Web搜索是Web信息发现的主要方式

搜索是最重要的Web应用



搜索是最重要的Web应用

- 没有搜索，很难找到所需的内容
- 没有搜索，在Web上创建内容也就缺了动机
 - 如果没人能够看到为什么要发布内容？
 - 如果没有任何回报为什么要发布内容？
- Web上必须要有人买单
 - 服务器、Web基础设施、内容创建过程等需要费用支持
 - 这些费用的大部分都是通过搜索广告支付
 - 可以说，搜索为Web买单

Web的特性

- **Web内容的产生机制**：无集中控制的无中心的网页内容发布机制，参与者的背景和动机具有空前的多样性
- **内容庞杂重复**：涉及数十种自然语言和数千种专业语言，风格存在巨大差异，充斥重复内容
- **可信度差**：可以无控制和无限制地发布内容，可能并不存在统一的、与用户无关的可信度标准
- **Web规模巨大，且不断增长**

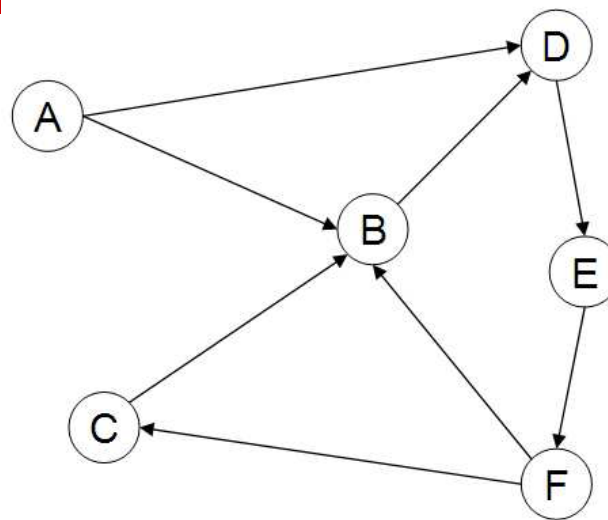
Web图

□ **Web图**：指以网页为节点，以网页之间的超链接为弧的有向图

□ **Web图的特性**

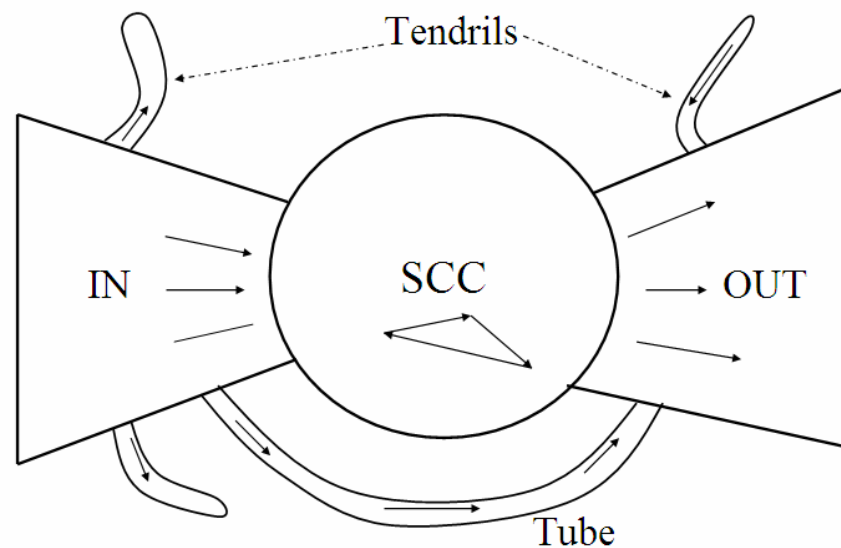
- 节点的平均入度8-15
- 超链接的分布满足幂分布定律：入度为 i 的网页总数目正比于 $1/i^\alpha$ ($\alpha = 2.1$)

□ 整个Web图呈蝴蝶结结构



Web 图

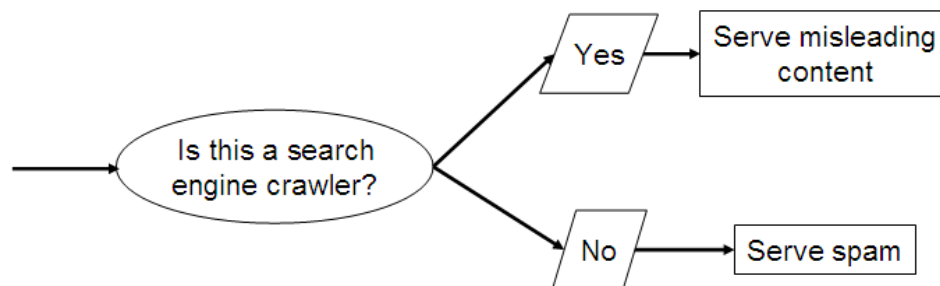
□ Web图的蝴蝶结结构



□ SCC: Strongly connected component

作弊网页(spam)

- **作弊网页**: 指采用针对性手段达到在搜索结果中得到较高排名目的的网页
- **根源**: 网页内容建设动机的多样性, 比如商业动机
- **几种方式**
 - **操作网页内容**: 比如故意重复关键词
 - **伪装(cloaking)**: 欺骗搜索引擎索引器



- **桥页**: 访问桥页时, 被重定向到另一网页

Web搜索的用户体验

□ Web搜索用户的特点

- Web搜索的用户数量巨大，不专业
- 用户查询需求种类多：信息类查询、导航类查询、事务类查询

□ Web搜索改进用户体验的举措

- 提高返回结果的相关性，即保证前面结果的正确率而不是召回率
- 用户体验轻量化，使查询页面和返回结果页面简洁整齐

Web搜索中的广告

- **源由：** Web运转需要经费支持
- Web搜索是一个很好的广告平台



第一代搜索广告: Goto (1996)

- ❑ **基本思想:** 预售查询, 搜索结果按照投标价格的顺序排序
- ❑ Buddy Blake 为此查询及搜索投出最高价 (\$0.38)
- ❑ 只要某个人点击了该链接, Buddy Blake就要付\$0.38的费用给 Goto 公司
- ❑ 不区分广告还是文档, 没有相关度排序, 仅仅是一个结果列表!
- ❑ ... 但是Goto并不假装存在相关度

第二代搜索广告: Google (2000)

- 严格区分一般搜索结果和广告搜索结果

第二代搜索广告: Google (2000)

Web Images Maps News Shopping Gmail more Sign in

Google discount broker Search Advanced Search Preferences

Web Results 1 - 10 of about 807,000 for discount broker [definition]. (0.12 seconds)

Discount Broker Reviews
Information on online discount brokers emphasizing rates, charges, and customer comments and complaints.
www.broker-reviews.us/ - 94k - Cached - Similar pages

Discount Broker Rankings (2008 Broker Survey) at SmartMoney.com
Discount Brokers. Rank/ Brokerage/ Minimum to Open Account, Comments, Standard Commission*, Reduced Commission, Account Fee Per Year (How to Avoid), Avg. ...
www.smartmoney.com/brokers/index.cfm?story=2004-discount-table - 121k - Cached - Similar pages

Stock Brokers | Discount Brokers | Online Brokers
Most Recommended. Top 5 Brokers headlines. 10. Don't Pay Your Broker for Free Funds May 15 at 3:39 PM. 5. Don't Discount the Discounters Apr 18 at 2:41 PM ...
www.fool.com/investing/brokers/index.aspx - 44k - Cached - Similar pages

Discount Broker
Discount Broker - Definition of Discount Broker on Investopedia - A stockbroker who carries out buy and sell orders at a reduced commission compared to a ...
www.investopedia.com/terms/d/discountbroker.asp - 31k - Cached - Similar pages

Discount Brokerage and Online Trading for Smart Stock Market ...
Online stock broker SogoTrade offers the best in discount brokerage investing. Get stock market quotes from this Internet stock trading company.
www.sgotrade.com/ - 39k - Cached - Similar pages

15 questions to ask discount brokers - MSN Money
Jan 11, 2004 ... If you're not big on hand-holding when it comes to investing, a discount broker can be an economical way to go. Just be sure to ask these ...
moneycentral.msn.com/content/Investing/StartInvesting/P66171.asp - 34k - Cached - Similar pages

Sponsored Links

Rated #1 Online Broker
No Minimums. No Inactivity Fee.
Transfer to Firsttrade for Free!
www.firsttrade.com

Discount Broker
Commission free trades for 30 days.
No maintenance fees. Sign up now.
TDAMERITRADE.com

TradeKing - Online Broker
\$4.95 per Trade, Market or Limit
SmartMoney Top Discount Broker 2001
www.TradeKing.com

Scottrade Brokerage
\$7 Trades, No Share Limit. In-Depth Research. Start Trading Online Now!
www.Scottrade.com

Stock trades \$1.99-\$3
100 free trades, up to \$100 back for transfer costs, \$500 minimum
www.sgotrade.com

\$3.95 Online Stock Trades
Market/Limit Orders, No Share Limit and No Inactivity Fees
www.Marsco.com

INGDIRECT | ShareBuilder
Discount Broker. No Inactivity Fee. No Account Fees.

SogoTrade出现在搜索结果中

SogoTrade出现在广告中

搜索引擎是不是把广告商的结果放在非广告商的结果之前?

所有的主流搜索引擎都否认这一点

广告排序方法

- **简单的方法:** 按照类似Goto的方式, 即按照对查询的投标价格进行排序
 - 可能造成相关性不好
- **改进方法:** 综合考虑投标价格和相关性进行排序
 - 相关度度量的关键指标: 点击率(clickthrough rate)
 - 结果: 无关的广告将得到很低的排名

Google次高竞标价格拍卖机制

advertiser	bid	CTR	ad rank	rank	paid
A	\$4.00	0.01	0.04	4	(minimum)
B	\$3.00	0.03	0.09	2	\$2.68
C	\$2.00	0.06	0.12	1	\$1.51
D	\$1.00	0.08	0.08	3	\$0.51

- **bid**: 每个广告商为每次点击给出的最大投标价格
- **CTR**: 点击率，即一旦被显示后被点击的比率
- **ad rank**: $\text{bid} \times \text{CTR}$: 这种做法可以在 (i) 广告商愿意支付的价钱 (ii) 广告的相关度高低 之间进行平衡。
- **rank**: 拍卖中的排名
- **paid**: 广告商的次高竞标价格

Google次高竞标价格拍卖机制

advertiser	bid	CTR	ad rank	rank	paid
A	\$4.00	0.01	0.04	4	(minimum)
B	\$3.00	0.03	0.09	2	\$2.68
C	\$2.00	0.06	0.12	1	\$1.51
D	\$1.00	0.08	0.08	3	\$0.51

- **次高竞标价格拍卖：** 广告商支付其维持在拍卖中排名所必须的价钱 (加上一分钱) (用它的下一名计算其支付价格)
- $\text{price}_1 \times \text{CTR}_1 = \text{bid}_2 \times \text{CTR}_2$
- $\text{price}_1 = \text{bid}_2 \times \text{CTR}_2 / \text{CTR}_1$
- $p_1 = \text{bid}_2 \times \text{CTR}_2 / \text{CTR}_1 = 3.00 \times 0.03 / 0.06 = 1.50$
- $p_2 = \text{bid}_3 \times \text{CTR}_3 / \text{CTR}_2 = 1.00 \times 0.08 / 0.03 = 2.67$
- $p_3 = \text{bid}_4 \times \text{CTR}_4 / \text{CTR}_3 = 4.00 \times 0.01 / 0.08 = 0.50$

具有高投标价格的关键词

参考<http://www.cwire.org/highest-paying-search-terms/>

\$69.1	mesothelioma treatment options
\$65.9	personal injury lawyer michigan
\$62.6	student loans consolidation
\$61.4	car accident attorney los angeles
\$59.4	online car insurance quotes
\$59.4	arizona dui lawyer
\$46.4	asbestos cancer
\$40.1	home equity line of credit
\$39.8	life insurance quotes
\$39.2	refinancing
\$38.7	equity line of credit
\$38.0	lasik eye surgery new york city
\$37.0	2nd mortgage
\$35.9	free car insurance quote

搜索广告的效益

- 每次用户点击广告，搜索引擎公司将会获得收益
- 用户只会点击其感兴趣的广告
 - 搜索引擎会对误导性和不相关的广告进行惩罚
 - 于是，用户在点击广告后往往会感到满意
- 广告商通过广告能够在物有所值的情况下找到新的客户

课堂思考

□ 为什么和TV、报纸和电台相比，Web搜索对广告商更有吸引力？

搜索广告的相关问题

□ 关键词套现(Keyword arbitrage)

- 比如从Google买一个关键词
- 然后将流量导向一个第三方页面，该页面对应机构付的钱将比你付给Google的多得多
- 该页面对于搜索用户来说基本没意义

□ 垃圾点击(click spam)

- 搜索用户对赞助搜索结果的非善意点击

搜索广告的相关问题

□ 商标侵权

- 例子：geico (美国政府雇员保险公司，是美国第四大私人客户汽车保险公司)
- 曾经搜索词项 “geico” 在Google上可以买到
- 导致Geico 在美国控告Google侵权
- Louis Vuitton(LV) 曾在欧洲控告Google侵权
- 参考 http://google.com/tm_complaint.html
- 如果采用商标做关键词，那么用户可能被误导到一个页面，该页面实际和用户期望购买的品牌产品无关

重复检测

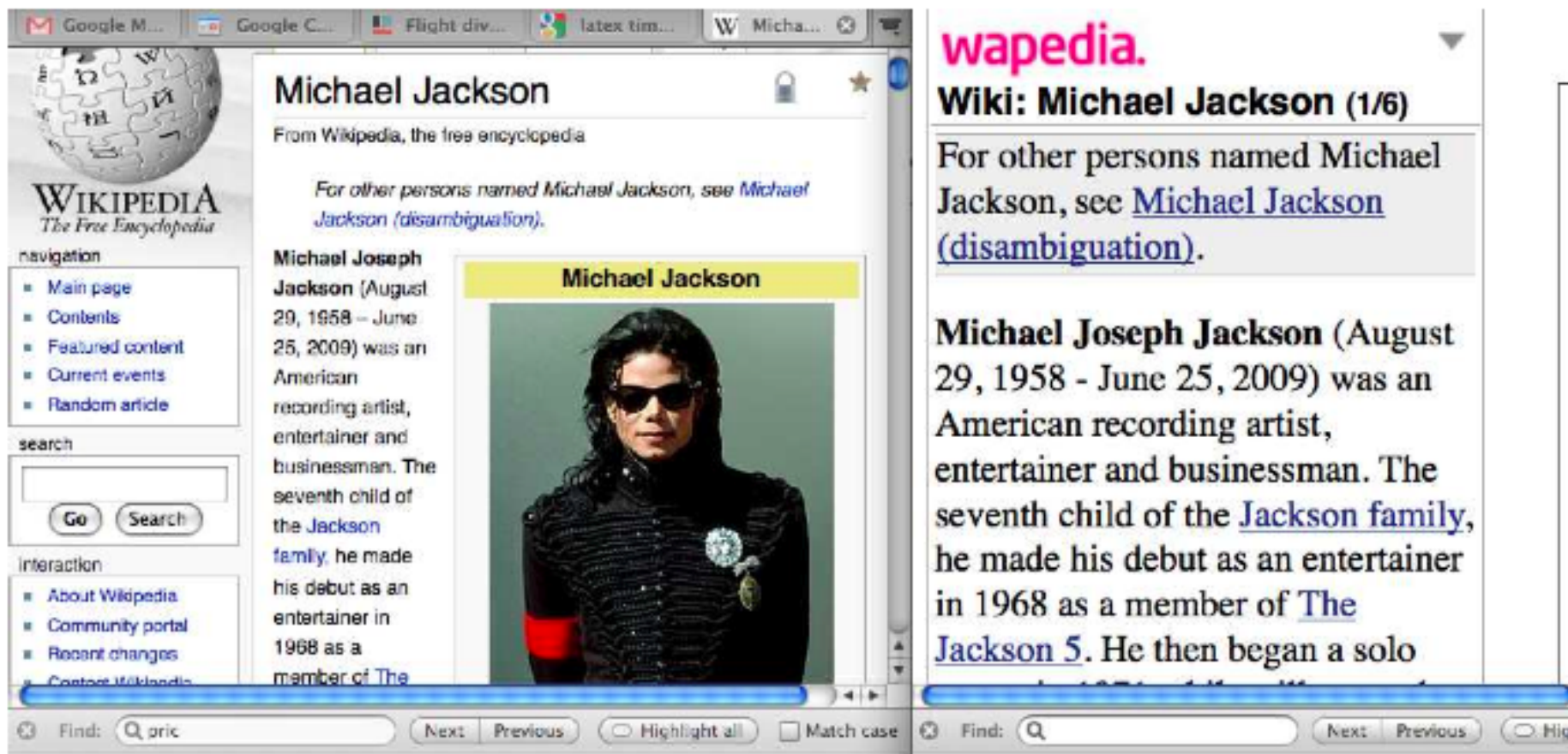
□ 必要性

- Web上充斥重复内容
- 对用户而言，如果搜索结果中存在不少几乎相同的页面，那么体验非常不好
- 去除重复可以降低存储和处理开销

□ 网页重复分类

- 完全重复(Exact duplicate)
 - 易剔除，比如采用哈希/指纹的方法
- 近似重复(Near-duplicate)
 - Web上存在大量近似重复，很难剔除

近似重复的例子



近似重复的检测

□ 基本思路

- 基于某种程度的内容相似度进行重复检测
- 基于“语法” (syntactic)而不是“语义” (semantic)进行页面的相似度评价
- 并不考虑那些内容意义上相似但是表达方式不同的近似重复

基于shingle集合的文档表示

- 每个 shingle 是一个基于词语的 k -gram
- 比如，对于 $k = 3$ ，那么文档 “a rose is a rose is a rose” 就可以表示成shingle的集合：
 $\{ \text{a-rose-is, rose-is-a, is-a-rose} \}$
- 可以基于哈希形式将shingle映射到 $1..2^m$ (例如 $m = 64$)之间，用 s_k 代表某个shingle映射到 $1..2^m$ 之间的一个数
- **将文档表示成shingle集合**，从而可以使用shingle来计算文档之间的语法相似度

Jaccard 系数

- 令 A 和 B 分别表示两个集合，则Jaccard系数为：

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

$(A \neq \emptyset \text{ or } B \neq \emptyset)$

- $\text{JACCARD}(A, A) = 1$
- $\text{JACCARD}(A, B) = 0$ if $A \cap B = \emptyset$
- Jaccard系数取值在 $[0, 1]$ 之间
- Jaccard系数刻画了两个集合的重合度
- 两个文档的相似度可以定义为它们的shingle集合之间的Jaccard距离

Jaccard 系数计算实例

□ 给定3篇文档：

d_1 : “Jack London traveled to Oakland”

d_2 : “Jack London traveled to the city of Oakland”

d_3 : “Jack traveled from Oakland to London”

□ 基于2-gram的shingle表示，可以计算它们之间的Jaccard距离如下：

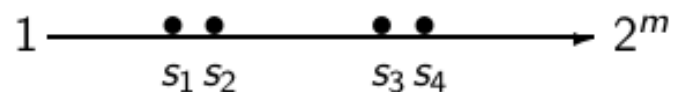
- $\mathcal{J}(d_1, d_2) = 3/8 = 0.375$

- $\mathcal{J}(d_1, d_3) = 0$

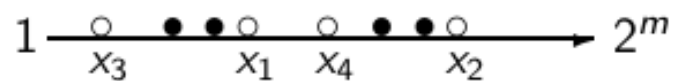
□ 可见，Jaccard系数对差异十分敏感

置换和最小值

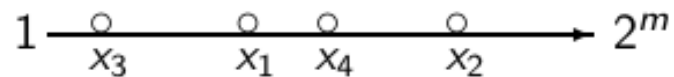
文档 1: $\{s_k\}$



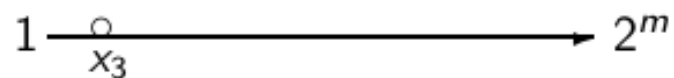
$$x_k = \pi(s_k)$$



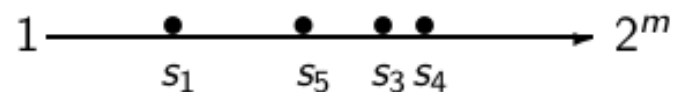
x_k



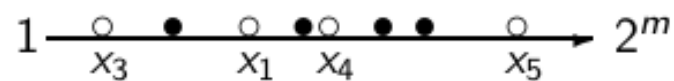
$$\min_{s_k} \pi(s_k)$$



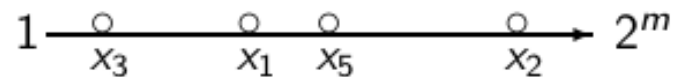
文档2: $\{s_k\}$



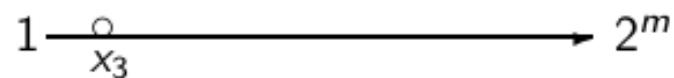
$$x_k = \pi(s_k)$$



x_k



$$\min_{s_k} \pi(s_k)$$



Jaccard系数的快速计算

- 将文档表示成梗概

$$\langle \min_{s \in d} \pi_1(s), \min_{s \in d} \pi_2(s), \dots, \min_{s \in d} \pi_{200}(s) \rangle$$

(一个200维的数字向量)

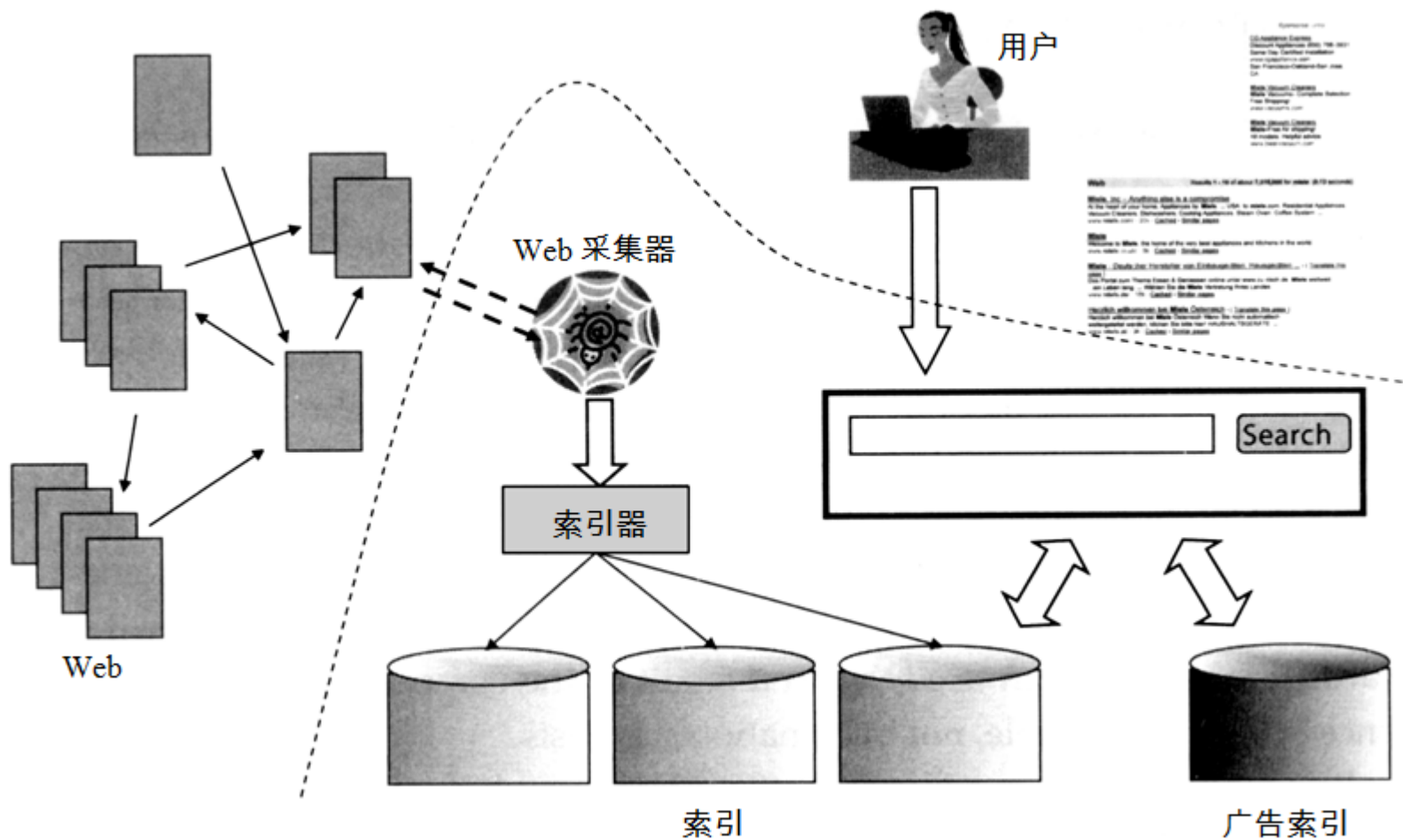
- 梗概是从文档shingle集合中精巧挑选出的子集

- 统计 $\langle d1, d2 \rangle$ 上的成功置换个数 k

- 置换 π 成功当且仅当 $\min_{s \in d1} \pi(s) = \min_{s \in d2} \pi(s)$

- 采用 $k/200$ 作为 $J(d1, d2)$ 的估计值

Web搜索系统组成



参考资料

□ 《信息检索导论》第 19 章

□ <http://ifnlp.org/ir>

□ Stanford 计算广告学课程,

<http://www.stanford.edu/class/msande239/>

课后作业

□ 见课程网页:

<http://10.76.3.31>