

现代信息检索

Modern Information Retrieval

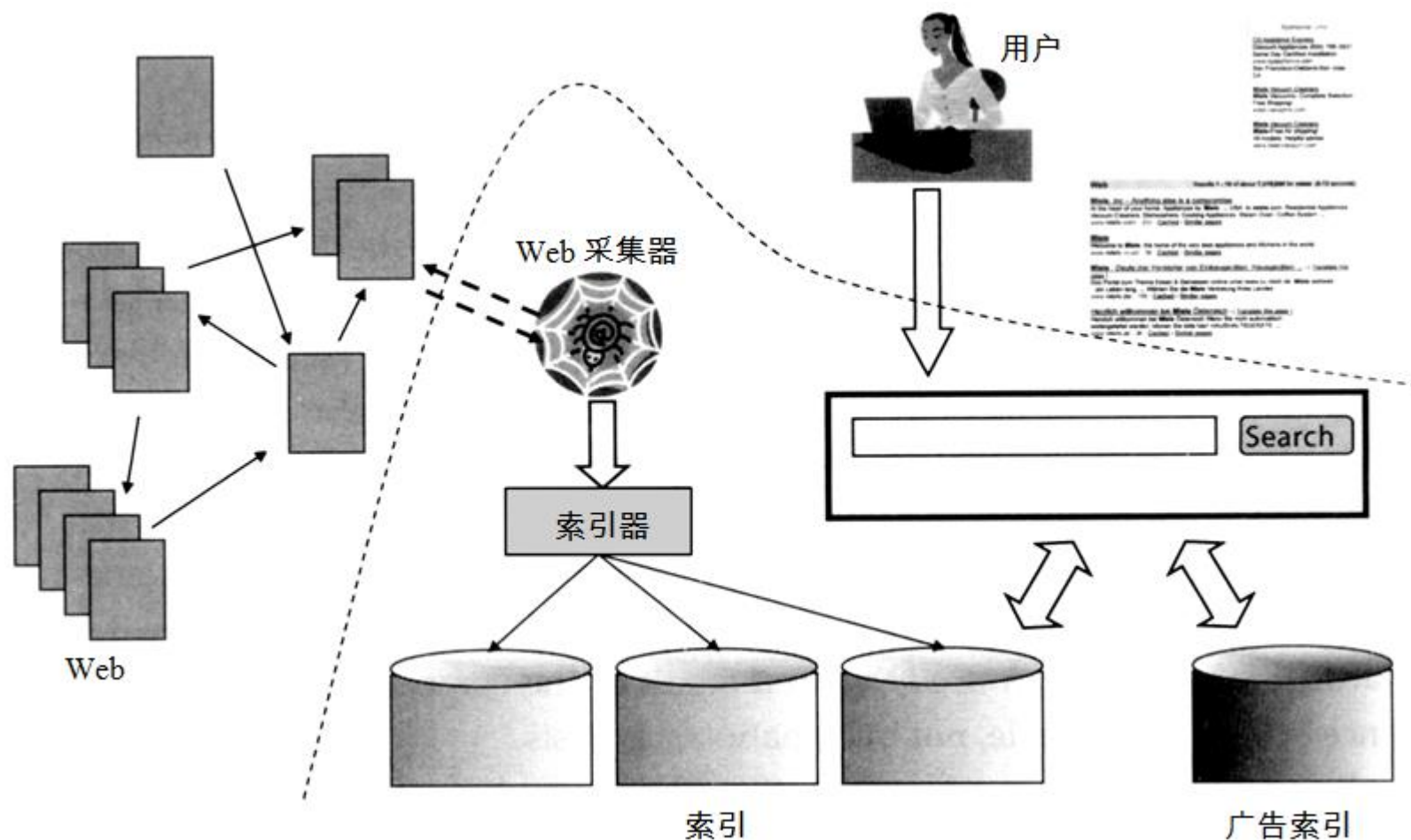
第14讲 Web搜索

Web Search

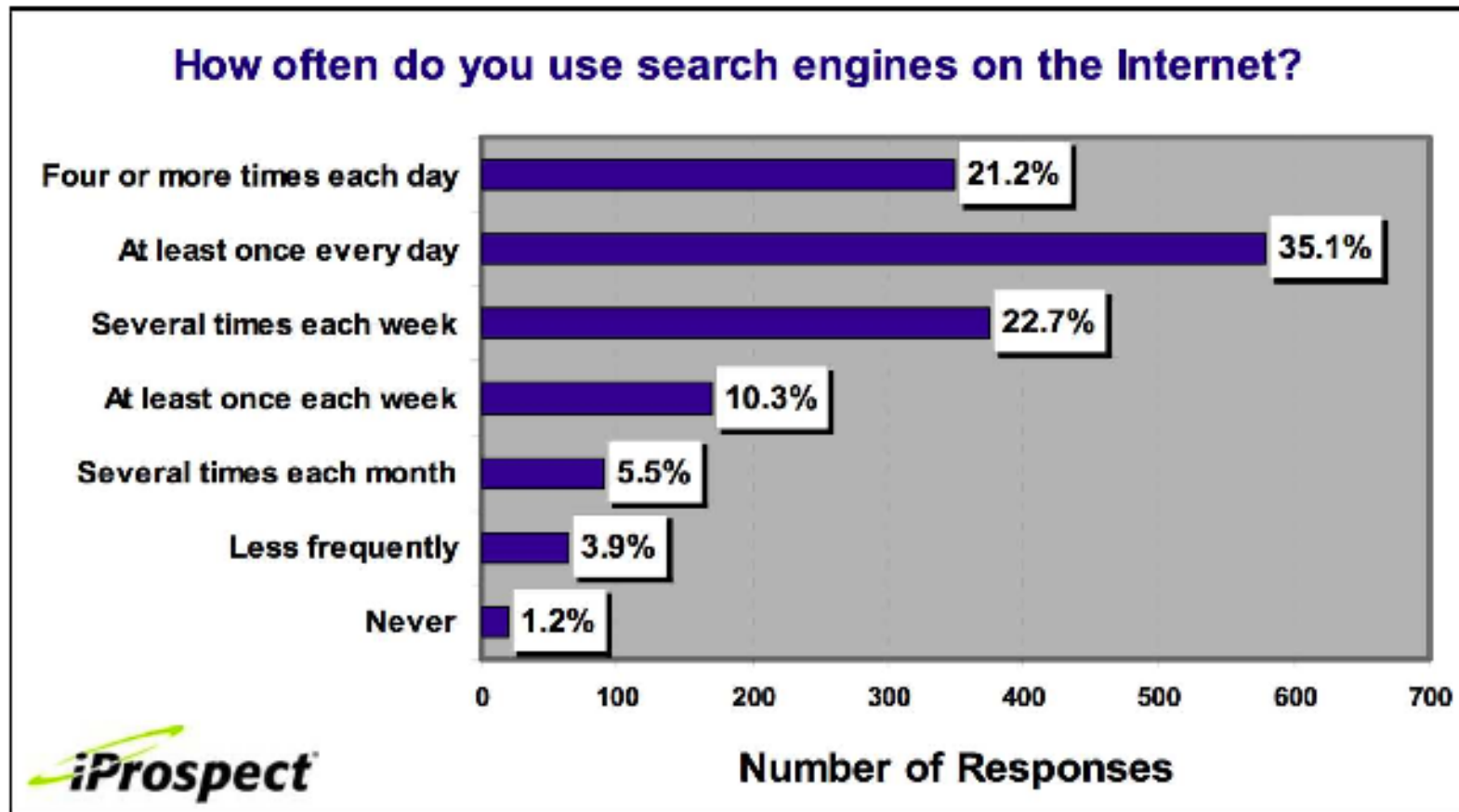
提纲

- 互联网上的搜索
- 互联网广告
- 重复检测
- 信息爬取

Web搜索系统组成



搜索是Web上使用最多的应用之一



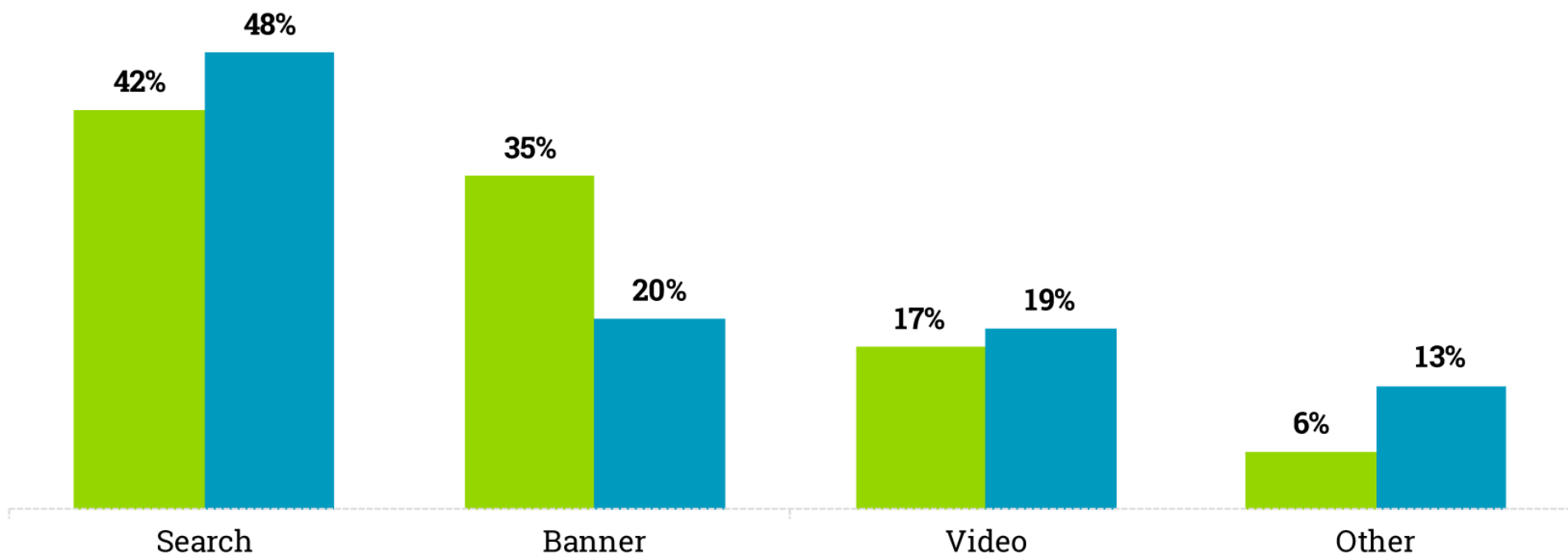
美国互联网广告利润统计

US Internet Advertising Revenue Share, by Channel in 2019



■ On mobile

■ On desktop



Published on MarketingCharts.com in June 2020 | Data Source: IAB / PwC

Based on IAB/PwC estimates / Read as: search accounted for 42% share of US mobile ad revenues in 2019

没有搜索引擎，Web甚至无法运转

- 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
 - 如果没人看为什么要发布内容？
 - 如果没有任何回报为什么要发布内容？
- Web运转必须要有人买单
 - 服务器、Web基础设施、内容创建过程等需要费用支持
 - 这些费用的相当大一部分都是通过搜索广告支付
 - 可以说，搜索为Web买单

兴趣聚合(Interest aggregation)

- Web的特点：具有相同兴趣的人，即使所处地理位置分散，也可以通过Web找到对方
 - 梅西/C罗/詹姆斯的球迷
 - 程序员 (开源项目和社区)
- 搜索引擎是实现兴趣聚合的关键事物

Web IR vs. 一般的IR

- 在Web上，搜索不仅仅是一个好的特点
 - 搜索是Web的关键事物: ...
 - ... 筹资、内容创建、兴趣聚合等等 → 参考搜索广告
- Web是一个充满噪声数据且组织失调的集合体 → 大量的重复需要检测
- 用户可以（某种意义上）无控制和无限制地发布内容 → 大量作弊内容需要检测

提纲

- 互联网上的搜索
- 互联网广告
- 重复检测
- 信息爬取

传统广告(1)

- 品牌广告(Brand Advertising)



传统广告(2)

■ 直接营销(Direct marketing)

Round Table PIZZA

PALO ALTO/EAST PALO ALTO
263 University Avenue
(Downtown/Delivery to Stanford)
650-322-2893

KING ARTHUR'S SUPREME

Round Table PIZZA

We Deliver

20% OFF Any Order
Offer excludes beverages, Manager's Specials, Kids Meal or any promotional items.
Offer valid on Drive-In, Carry-out, or Delivery. Limited delivery areas & hours. Weekend delivery fee may apply. Not valid with any other offer or discounts. Expires 7/1/06.

\$15.99 OFF Any Large Specialty Pizza
Original or Skinny Crust only.
Offer valid on Drive-In, Carry-out, or Delivery. Limited delivery areas & hours. Weekend delivery fee may apply. Not valid with any other offer or discounts. Expires 7/1/06.

\$11.99 OFF Any Large 1-Topping Pizza
Original or Skinny Crust only.
Offer valid on Drive-In, Carry-out, or Delivery. Limited delivery areas & hours. Weekend delivery fee may apply. Not valid with any other offer or discounts. Expires 7/1/06.

\$12.99 OFF Any Large 2-Topping Pizza
Original or Skinny Crust only.
Offer valid on Drive-In, Carry-out, or Delivery. Limited delivery areas & hours. Weekend delivery fee may apply. Not valid with any other offer or discounts. Expires 7/1/06.

\$5.00 OFF ANY X-LARGE PIZZA
\$4.00 OFF ANY LARGE PIZZA
\$2.00 OFF ANY MEDIUM PIZZA
Offer valid on Drive-In, Carry-out, or Delivery. Limited delivery areas & hours. Weekend delivery fee may apply. Not valid with any other offer or discounts. Expires 7/1/06.

FREE Medium 1 Topping Pizza
with the purchase of any Large or X-Large Specialty at regular menu price.
Offer valid on Drive-In, Carry-out, or Delivery. Limited delivery areas & hours. Weekend delivery fee may apply. Not valid with any other offer or discounts. Expires 7/1/06.

ADVERTISEMENT WITH MONEY MAILER OF PALO ALTO/LOS ALTOS/STANFORD (550) 955-1238
226-07-0082 226-07-0082

H.O.T! Coupons Web Ad • 226-07-0082F

©2006 Money Mailer LLC
<http://www.hotcoupons.com>

传统广告的不足

- 广告投放场地或媒介相对有限：报纸、电视、杂志、橱窗、公汽、电梯等
 - 广告场地的费用一般不菲：CCTV 标王
 - 很难进行个性化
 - 投放效果取决于广告商的智慧
 - 投放效果很难度量
-
- 互联网的出现改变了这一切.....

互联网广告的优点

- 无限机会
- 无限创意
- 完全可以个性化处理
- 每次点击花费的代价很低
- 定量度量程度高

互联网广告的主要形式(1)

■ 图片广告



图片广告

互联网广告的主要形式(2)

- 文本广告
 - 搜索关键词触发的广告(Sponsored Search driven Ad, paid Ad), 也称搜索广告。Google Adwords
 - 网页内容触发的广告(Web Page driven Ad), 也称上下文广告(Contextual Ad)。Google Adverb

互联网文本广告的主要类型(1)

■ 搜索广告

The image shows a screenshot of a Google search results page for the query "andrei broder". The search bar at the top contains the text "andrei broder" and a "Search" button. To the right of the search bar are links for "Advanced Search" and "Preferences". Below the search bar, the text "Web" is on the left, and "Results 1 - 10 of about 90,100 for andrei broder. (0.12 seconds)" is on the right. The main content area lists several search results. On the right side of the page, there is a section titled "Sponsored Links" which is circled in yellow. This section contains three advertisements: "Borders Official Site", "Yahoo! is Hiring", and "Yahoo! Search Blog: Search in the Future".

Google andrei broder Search Advanced Search Preferences

Web Results 1 - 10 of about 90,100 for andrei broder. (0.12 seconds)

Andrei Broder - Wikipedia, the free encyclopedia
Andrei Zary Broder (Hebrew: אנדרזי ברודר) is a Research Fellow and Vice President of Emerging Search Technology for Yahoo!. He previously has worked for ...
en.wikipedia.org/wiki/Andrei_Broder - 21k - [Cached](#) - [Similar pages](#) - [Note this](#) - [Filter](#)

Andrei Broder | Yahoo! Research
Andrei Broder is a Yahoo! Research Fellow and Vice President for Computational Advertising. Previously he was an IBM Distinguished Engineer and the CTO of ...
research.yahoo.com/Andrei_Broder - 43k - [Cached](#) - [Similar pages](#) - [Note this](#) - [Filter](#)

DBLP: Andrei Z. Broder
63, Andrei Z. Broder: Introduction: The Fourth International Workshop on ... 4, Andrei Z. Broder: A Provably Secure Polynomial Approximation Scheme for the ...
www.informatik.uni-trier.de/~ley/db/indices/a-tree/b/Broder,Andrei_Z.html - 79k - [Cached](#) - [Similar pages](#) - [Note this](#) - [Filter](#)

Andrei Broder Joins Yahoo - Search Marketing News Blog - Search ...
Nov 18, 2005 ... Andrei Broder, former vice president of research at AltaVista and until recently Distinguished Engineer & CTO, IBM Research, ...
blog.searchenginewatch.com/blog/061118-122644 - 33k - [Cached](#) - [Similar pages](#) - [Note this](#) - [Filter](#)

Yahoo! Search Blog: Search in the Future
Feb 1, 2006 ... Questions for Andrei Broder re: emerging search technology? ... Yesterday evening, we conducted the interview with Andrei Broder and we ...
www.ysearchblog.com/archives/000242.html - 45k - [Cached](#) - [Similar pages](#) - [Note this](#) - [Filter](#)

Sponsored Links

Borders Official Site
Shop Borders For Books, Music, DVDs, Gifts & More In One Online Store
www.borders.com

Yahoo! is Hiring
Work For One of the Best. Start Your Career with Yahoo! Today.
careers.yahoo.com
San Francisco-Oakland-San Jose, CA

■ 网页广告(内容匹配广告, Content Match)

[HOME](#)
[U.S.](#)
[BUSINESS](#)
[WORLD](#)
[ENTERTAINMENT](#)
[SPORTS](#)
[TECH](#)
[POLITICS](#)
[ELECTIONS](#)

[Entertainment Video](#)
[Celebrity](#)
[TV](#)
[Movies](#)
[Music](#)
[Reviews](#)
[Fashion](#)
[Books](#)
[Arts](#)

Beastie Boys recording "political" album

By John Benson — Fri Oct 24, 4:13 am ET

Buzz Up
 Send
 Share
 Print

CLEVELAND (Billboard) – As the Beastie Boys kick off their Get Out and Vote tour, the group is also working on its 2007 instrumental album “The Mixtape.”

“We’re actually in the middle of it,” says Adam “Ad-Rock” Horowitz to Reuters. “It’s due sometime next year. It’s a lot of different styles. And it’s political, depending on what you want to hear. Toilet talk and fart jokes are yeah, very.”

Any chance of new material from the group’s 2008 tour? “I don’t think so,” he says. “If you play the new songs that we’ve been writing, it always seems like people want to hear them.”

Reuters — Michael ‘Mike D’ Diamond (L) and Adam ‘MCA’ Yauch of the Beastie Boys perform on the main stage during ...

Video: D.L. Hughley on the election CNN

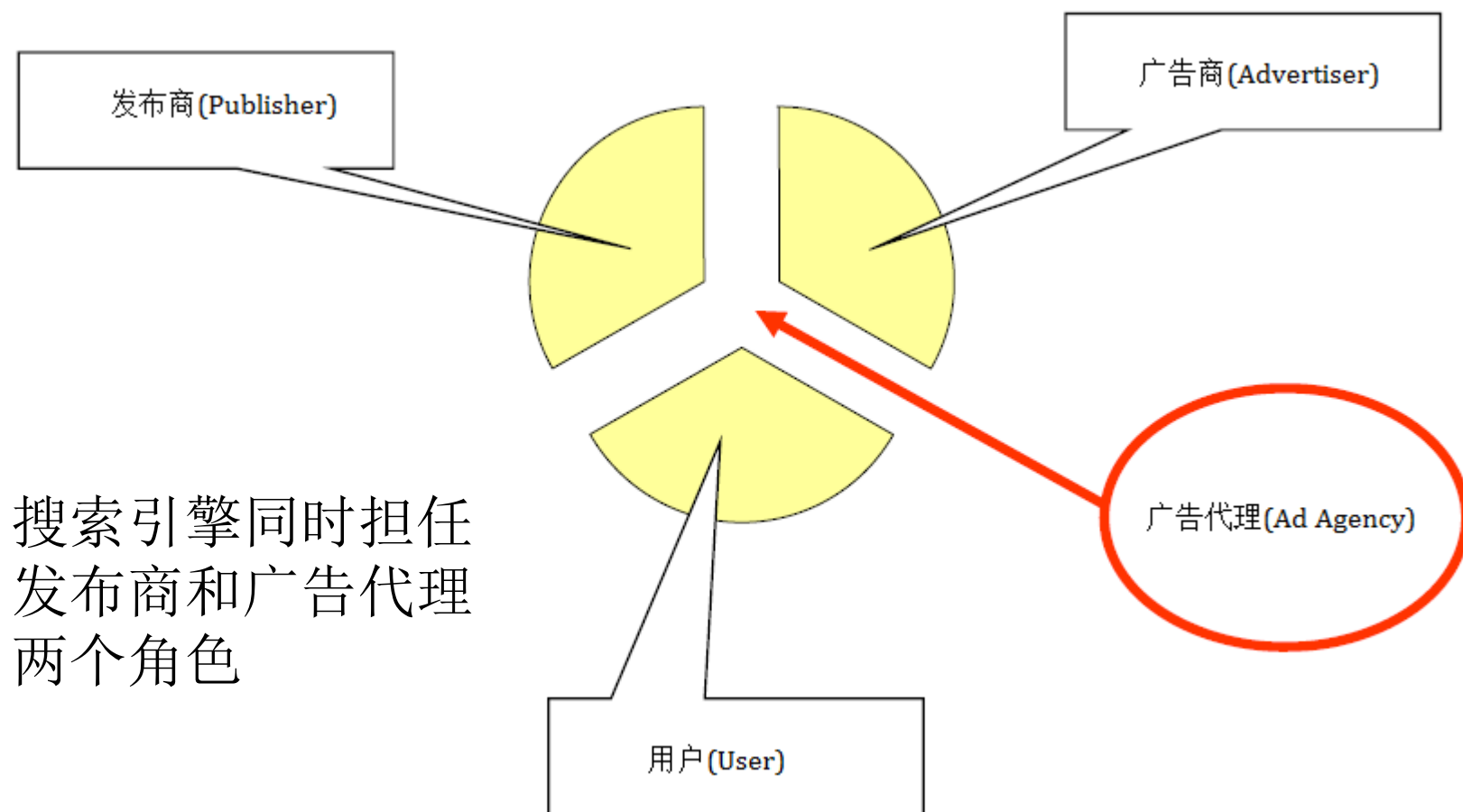
sponsored links

[Beastie Boys Vs. Jay-Z - New York](#)
Beastie Boys battle Jigga for NY supremacy. Check it out on fuse.
[fuse.tv/all-american-face-off](#)

[Beastie Boys](#)
Browse a huge selection now. Find exactly what you want today.
[www.ebay.com](#)

[Cell Phone Ringtones](#)
Download Your Favorite Ringtone In Seconds.
[www.RingtoneOcean.com](#)

互联网广告中的利益三方



第一代搜索广告: Goto (1996) 竞价排名

The screenshot shows a search results page from Goto.com. The search query is 'Wilmington real estate'. The page features a yellow sidebar on the left and a main content area with a yellow header. A yellow box in the main area promotes 'Access 75% of all users now!'. Below this, there are three sponsored listings, each with a bid price circled in red.

Wilmington real estate.

Access 75% of all users now!
Premium Listings reach 75% of all
Internet users. [Sign up](#) for Premium
Listings today!

1. [Wilmington Real Estate - Buddy Blake](#)
Wilmington's information and real estate guide. This is your on
anything to do with Wilmington.
[www.buddyblake.com](#) (Cost to advertiser: **\$10.28**)
2. [Coldwell Banker Sea Coast Realty](#)
Wilmington's number one real estate company.
[www.cbseacoast.com](#) (Cost to advertiser: **\$10.37**)
3. [Wilmington, NC Real Estate Becky Bullard](#)
Everything you need to know about buying or selling a home c
on my Web site!
[www.iwwc.net](#) (Cost to advertiser: **\$10.35**)

第一代搜索广告: Goto (1996)



- Buddy Blake 为此搜索投出最高价 (\$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
- 搜索结果按照投标价格的顺序排序 – Goto可以获得最大的收益
- 不区分广告还是文档, 仅仅是一个结果列表!
- 广告预售, 坦诚公开, 没有相关度排序
- ... 但是Goto并不假装存在相关度

第二代搜索广告: Google (2000/2001)

- 严格区分搜索结果和搜索广告

两个列表结果: web 网页 (左图) 及广告 (右图)

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 (2003 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.foo.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.sogotrade.com/](#) - 38k - 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/P68171.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 \$1.95-\$3
100 free trades, up to \$100 back for transfer costs, \$500 minimum
[www.sogotrade.com](#)

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

INGDIRECT | ShareBuilder
Discount Broker, Online Broker, Stock Broker, ...

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

广告是否会影响编辑的内容？

- 在报纸、电视上存在类似问题
- 报纸一般不会刊登针对其主要广告商的严厉指责性质的文章
- 在报纸和TV上，广告和编辑内容之间的界限往往变得很模糊
- 现在还不清楚搜索引擎广告是否和上面一样

广告在右部如何排序？

Web Images Maps News Shopping Gmail more

Sign in



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.sogotrade.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** 2007

www.TradeKing.com

Scottrade Brokerage

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

www.Scottrade.com

Stock trades \$1.50 - \$3

100 free trades, up to \$100 back
for transfer costs, \$500 minimum

www.sogotrade.com

\$3.95 Online Stock Trades

Market/Limit Orders, No Share Limit
and No Inactivity Fees

www.Marsco.com

INGDIRECT | ShareBuilder

如何对广告排序？

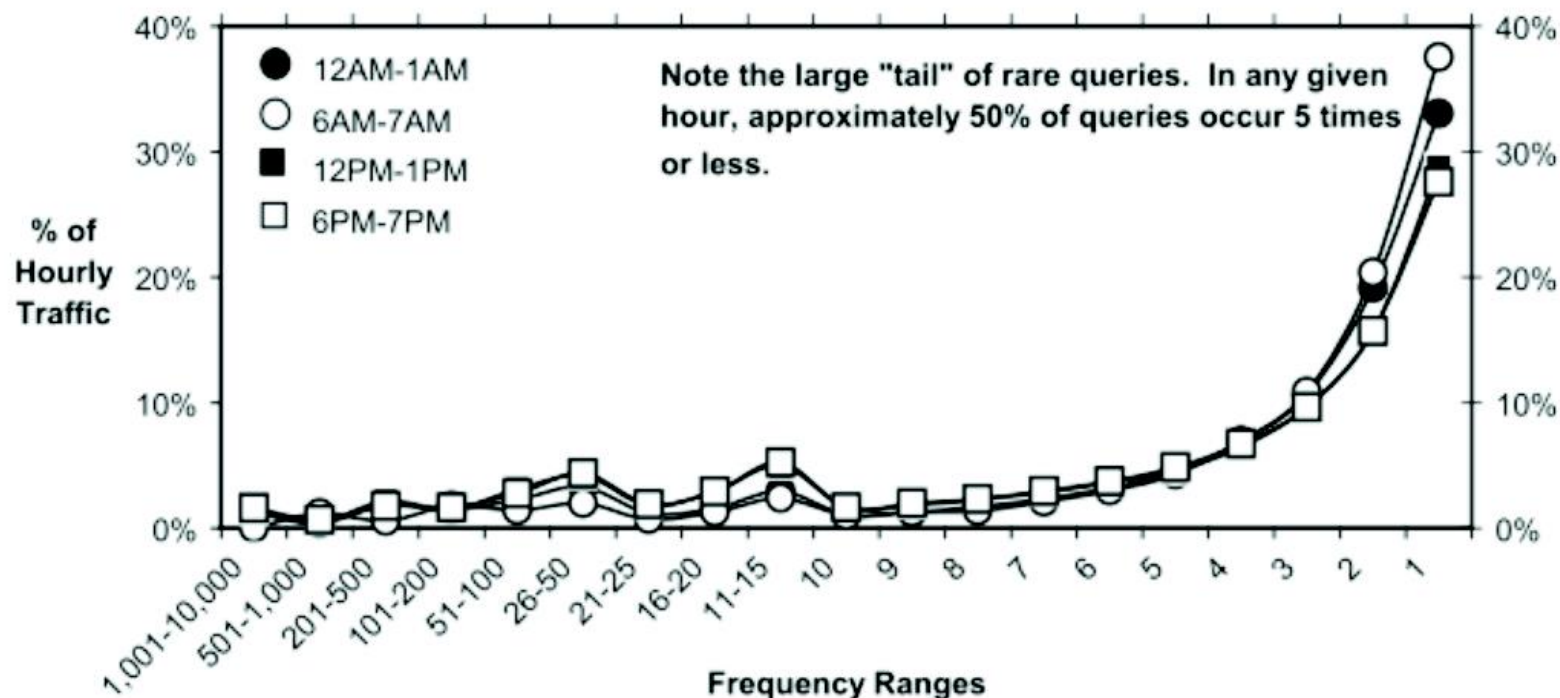
- 广告商对关键词竞标 – 拍卖方式
- 拍卖系统公开，任何人都可以参与关键词竞标
- 广告商仅在用户点击广告商才真正付费
- 拍卖机制如何确定某条广告的排序以及该广告的支付价格？
- 基本思路是次高价拍卖(second price auction)原则
- 搜索引擎中最重要的研究领域之一---计算广告学
 - 对每条广告压榨出一分钱也就意味着为搜索引擎公司带来上亿的额外收益

如何对广告排序？

- 简单的方法: 按照类似Goto的方式, 即按照投标价格排序
 - 不好的方法: 可能会被滥用
 - 例如: query [does my husband cheat?] → 有关离婚律师的广告
 - 我们并不想得到与用户查询不相关的广告
- 替代方法: 按照投标价格和相关性排序
- 相关度度量的关键指标: 点击率 (clickthrough rate, CTR)
 - $CTR = \text{clicks per impressions}$
- 结果: 无关的广告将得到很低的排名
 - 即使在短期时间内降低了搜索引擎的收益
 - 希望: 如果用户能通过系统获得有价值的信息, 那么系统的总体接受程度和整体收益将最大化
- 其他排序因子: 位置、一天内的时间、着陆页(landing page)的质量和装载速度
- 最主要的排序因子: 查询

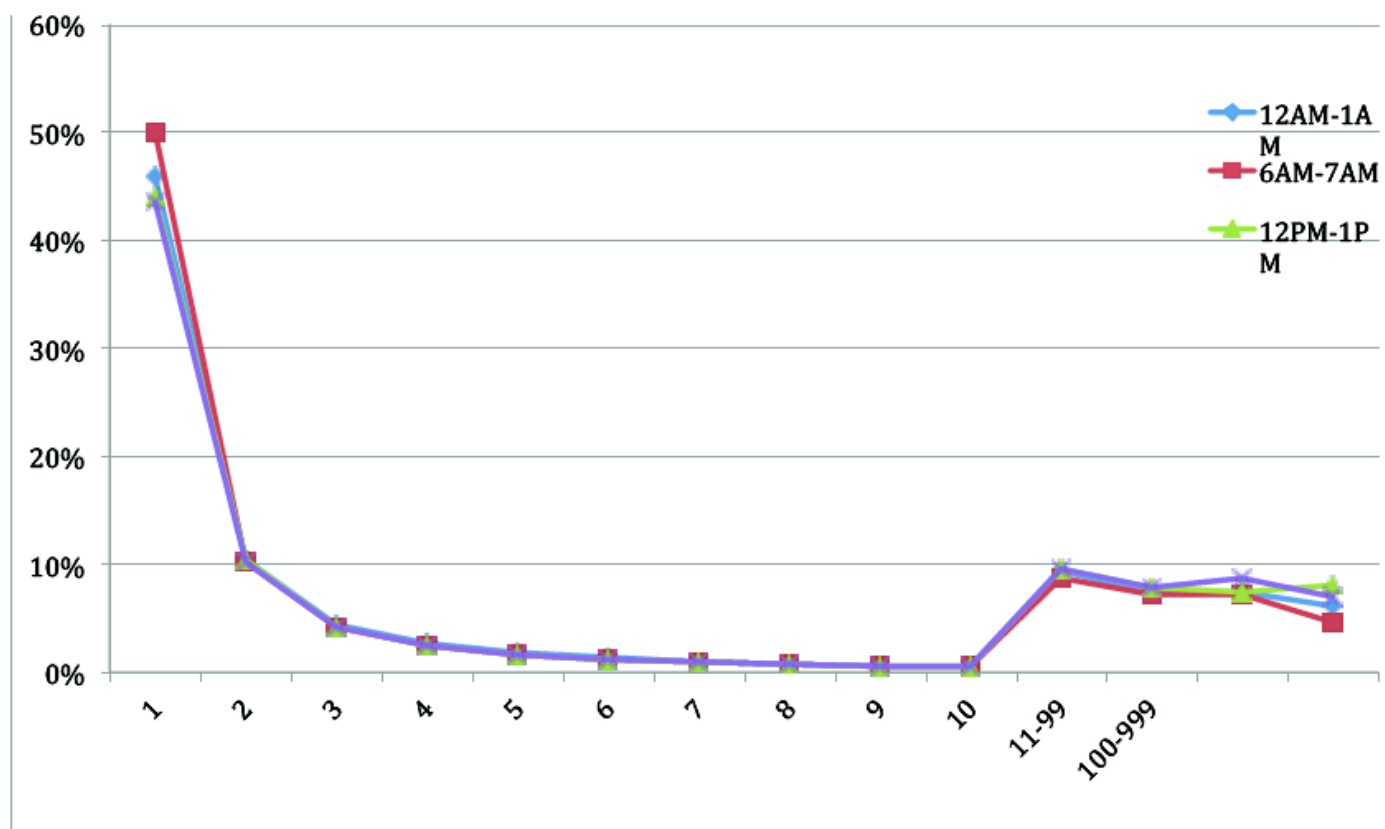
Web查询“长尾”现象：基于AOL查询频次的统计

- 罕见查询形成一个大“尾巴”。大约50%的查询的频次小于等于5。



基于查询频次的流量统计（数据来源：Yahoo）

- 大部分流量都在“头”（常见）和“尾”（罕见）



长尾效应的解释

- 有两种解释：
 - 大多数用户搜索“常见”查询；一小部分用户搜索“罕见”查询
 - 大量用户使用“常见”查询；同时大量用户也会使用一些“罕见”查询
- 对在线零售的研究支持第二种解释[Goel et al CIKM 2009]
 - 每个人都有一些不同常人的独特喜好，所以热销或冷门商品都会购买
 - 但是，不同顾客的独特程度不同
 - 对一站式营销的研究表明，罕见商品的存在可以提升常见商品的销量
 - 对于搜索引擎，对罕见查询的有效覆盖可以提升用户满意度，可望在将来从常见查询获益

Google AdsWords 的例子

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。CTR是一种相关性度量指标。
- **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（基于ad rank）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)

$$\text{price}_1 \times \text{CTR}_1 = \text{bid}_2 \times \text{CTR}_2 \text{ (使得排名 } \text{rank}_1 = \text{rank}_2 \text{)}$$

$$\text{price}_1 = \text{bid}_2 \times \text{CTR}_2 / \text{CTR}_1$$

$$p_1 = \text{bid}_2 \times \text{CTR}_2 / \text{CTR}_1 = 3.00 \times 0.03 / 0.06 = 1.50$$

$$p_2 = \text{bid}_3 \times \text{CTR}_3 / \text{CTR}_2 = 1.00 \times 0.08 / 0.03 = 2.67$$

$$p_3 = \text{bid}_4 \times \text{CTR}_4 / \text{CTR}_3 = 4.00 \times 0.01 / 0.08 = 0.50$$

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：三赢？

- 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

课堂练习

- 为什么和TV、报纸和电台相比，Web搜索对广告商更有吸引力？
- 广告商会为所有一切买单，那么它们会受到欺骗吗？
- 这对用户来说究竟是好消息还是坏消息？
- 当然，不论如何做，这都会危害搜索引擎

并非三赢：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
 - 比如，重定向到一个满是广告的页面
- 该页面对于搜索用户来说基本没意义
- 广告作弊者一直在钻营想招
- 搜索引擎必须要要花时间对付他们

并非三赢：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)
- 2005年的部分时间内: 搜索词项 “geico” 在Google上可以买到
- Geico 在美国控告Google侵权
- Louis Vuitton(LV) 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


提纲

- 互联网上的搜索
- 互联网广告
- 重复检测
- 信息爬取

重复检测

- Web上充斥重复内容
- 相对其它文档集合，Web上的重复内容更多
- **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
- 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
- 边缘相关度(Marginal relevance)为0：如果一篇高度相关的文档出现在另一篇高度近似的文档之后，那么该文档变得不相关(冗余)
- 必须要去除这些近似重复

近似重复的例子



WIKIPEDIA
The Free Encyclopedia

navigation

- Main page
- Contents
- Featured content
- Current events
- Random article

search

Go Search

interaction

- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia


Michael Jackson

From Wikipedia, the free encyclopedia

For other persons named Michael Jackson, see [Michael Jackson \(disambiguation\)](#).

Michael Joseph Jackson (August 29, 1958 – June 25, 2009) was an American recording artist, entertainer and businessman. The seventh child of the [Jackson family](#), he made his debut as an entertainer in 1968 as a member of The

Michael Jackson



wapedia.

Wiki: Michael Jackson (1/6)

For other persons named Michael Jackson, see [Michael Jackson \(disambiguation\)](#).

Michael Joseph Jackson (August 29, 1958 - June 25, 2009) was an American recording artist, entertainer and businessman. The seventh child of the [Jackson family](#), he made his debut as an entertainer in 1968 as a member of [The Jackson 5](#). He then began a solo

课堂思考题

如何去掉Web上的近似重复页面呢？

近似重复的检测

- 采用编辑距离指标计算页面之间的相似度
- 需要说明的是，我们希望检测那些“语法”(syntactic)上而不是“语义”(semantic)上相似的页面
 - 页面的语义相似度(即内容语义之间的相似度)非常难以计算
- 也就是说，我们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
- 引入一个相似度阈值 θ 来判定“两个页面之间是否近似重复”
- 比如，如果两篇文档的相似度 $>$ 阈值 $\theta = 80\%$ ，那么认为两篇文档近似重复

将每篇文档表示成一个shingle集合

- 每个 shingle 是一个基于词语的n-gram
- 使用shingle来计算文档之间的语法相似度
- 比如，对于 $n = 3$ ，那么文档“a rose is a rose”就可以表示成shingle的集合：
 - { a-rose-is, rose-is-a, is-a-rose }
- 我们可以通过指纹(fingerprinting)算法将shingle映射到 $1..2^m$ (例如 $m = 64$)之间
- 接下来我们用 s_k 来代表某个shingle映射到 $1..2^m$ 之间的一个数
- 两个文档的相似度定义为它们的shingle集合的Jaccard距离

Jaccard距离计算回顾

- 一个常用的计算两个集合重合度的方法
- 令 A 和 B 分别表示两个集合
- Jaccard距离:

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

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

- $\text{JACCARD}(A, A) = 1$
- $\text{JACCARD}(A, B) = 0$ if $A \cap B = \emptyset$
- 并不要求 A 和 B 的大小一样
- Jaccard距离取值在 $[0, 1]$ 之间

Jaccard距离计算的例子

- 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，可以计算文档之间的Jaccard距离如下：
 - $J(d_1, d_2) = 3/8 = 0.375$
 - $J(d_1, d_3) = 0$
- 注意：Jaccard距离对差异十分敏感

将文档表示成梗概(sketch)

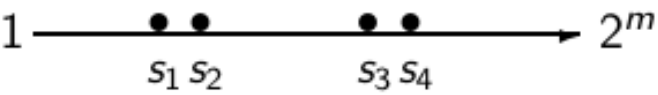
- 每篇文档的shingle的个数非常大
- 为提高效率，接下来我们使用文档的梗概来表示文档，它由文档的shingle集合中精巧挑选出的子集构成
- 比如，梗概中shingle的数目为 $n = 200 \dots$
- \dots 通过一系列置换 $\pi_1 \dots \pi_{200}$ 来定义
- 每个置换 π_i 都是 $1..2^m$ 上的随机置换
- 文档 d 的梗概定义为：
$$\langle \min_{s \in d} \pi_1(s), \min_{s \in d} \pi_2(s), \dots, \min_{s \in d} \pi_{200}(s) \rangle$$

(一个200维的数字向量)

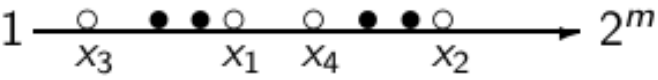
注：“置换”的英文原文是Permutation, 也可以译为“排列”

置换和最小值：例子

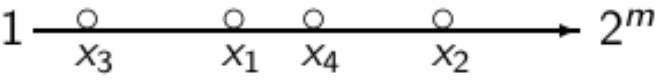
文档 1: $\{s_k\}$



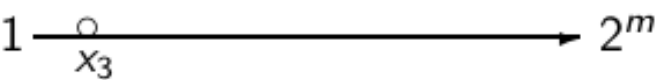
$$x_k = \pi(s_k)$$



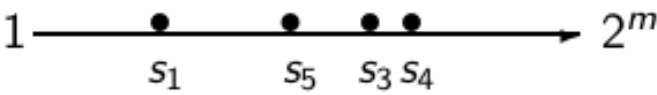
x_k



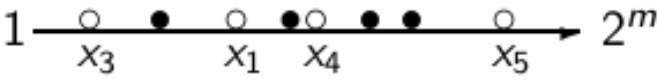
$$\min_{s_k} \pi(s_k)$$



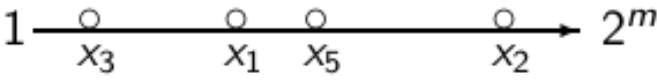
文档2: $\{s_k\}$



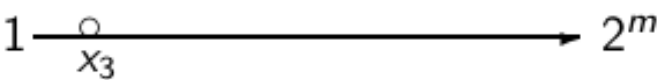
$$x_k = \pi(s_k)$$



x_k



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



使用 $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$ 作为文档 d_1 和 d_2 是否近似重复的测试条件。该例子中置换 π 表明: $d_1 \approx d_2$

计算梗概之间的Jaccard距离 (1)

- 现在每篇文档都变成了一个 $n=200$ 维的数字向量
- 该向量比高维空间下的shingle容易处理得多
- 如何计算Jaccard距离?

计算梗概之间的Jaccard距离 (2)

- 如何计算?
- 令 U 为 d_1 和 d_2 的并集, I 为它们的交集
- 对于 U 而言就存在 $|U|!$ 个置换
- 对于 $s' \in I$, 有多少置换 π 会使得
 $\operatorname{argmin}_{s \in d_1} \pi(s) = s' = \operatorname{argmin}_{s \in d_2} \pi(s)$?
- 答案是: $(|U| - 1)!$
- 对于 I 的每个 s , 存在着 $(|U| - 1)!$ 个不同的置换集合 \Rightarrow
 于是总共有 $|I|(|U| - 1)!$ 个置换能够保证
 $\operatorname{argmin}_{s \in d_1} \pi(s) = \operatorname{argmin}_{s \in d_2} \pi(s)$ 为真
- 因此, 使得 $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$ 为真的置换比例为:

$$\frac{|I|(|U| - 1)!}{|U|!} = \frac{|I|}{|U|} = J(d_1, d_2)$$

Jaccard距离估计

- 因此，成功的置换比例就等于Jaccard距离
 - 置换 π 成功当且仅当 $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$
- 随机选取一个置换，当置换成功是输出1，否则输出0，该过程是一个贝努利试验过程
- 成功概率的估计：在 n 次贝努利试验中成功比率 ($n = 200$)
- 我们使用的梗概基于置换的随机选择
- 因此，为了计算Jaccard距离，统计 $\langle d_1, d_2 \rangle$ 上的成功置换个数 k ，然后除以 $n = 200$.
- $k/n = k/200$ 就是 $J(d_1, d_2)$ 的估计值

实现

- 使用哈希函数来实现高效的置换:
$$h_i : \{1..2^m\} \rightarrow \{1..2^m\}$$
- 以任意顺序扫描两个集合并集中的所有shingle s_k
- 对每个哈希函数 h_i 及文档 d_1, d_2, \dots : 在某个固定存储位置（即下一页例子中的slot）中保留当前的最小值
- 如果 $h_i(s_k)$ 小于当前的最小值，那么对固定存储位置上的值进行更新

例子

定义两个哈希函数 $h(x)$ ， $g(x)$ 。表格中自上而下表示每一步对 s_i 的计算，slot表示当前最小值

	d_1	d_2
s_1	1	0
s_2	0	1
s_3	1	1
s_4	1	0
s_5	0	1
$h(x) = x \bmod 5$		
$g(x) = (2x + 1) \bmod 5$		
$\min(h(d_1)) = 1 \neq 0 =$		
$\min(h(d_2)) \quad \min(g(d_1)) =$		
$2 \neq 0 = \min(g(d_2))$		

$$\hat{J}(d_1, d_2) = \frac{0+0}{2} = 0$$

	d_1 slot		d_2 slot	
h	∞		∞	
g	∞		∞	
$h(1) = 1$	1	1	–	∞
$g(1) = 3$	3	3	–	∞
$h(2) = 2$	–	1	2	2
$g(2) = 0$	–	3	0	0
$h(3) = 3$	3	1	3	2
$g(3) = 2$	2	2	2	0
$h(4) = 4$	4	1	–	2
$g(4) = 4$	4	2	–	0
$h(5) = 0$	–	1	0	0
$g(5) = 1$	–	2	1	0

最终的梗概

课堂练习

	d_1	d_2	d_3
s_1	0	1	1
s_2	1	0	1
s_3	0	1	0
s_4	1	0	0

$h(x) = 5x + 5 \bmod 4$ Estimate $\hat{J}(d_1, d_2)$,
 $g(x) = (3x + 1) \bmod 4$

$\hat{J}(d_1, d_3), \hat{J}(d_2, d_3)$

解答 (1)

	d_1	d_2	d_3
s_1	0	1	1
s_2	1	0	1
s_3	0	1	0
s_4	1	0	0

$h(x) = (5x + 5) \bmod 4$
 $g(x) = (3x + 1) \bmod 4$

	d_1 slot	d_2 slot	d_3 slot
	∞	∞	∞
	∞	∞	∞
$h(1) = 2$	— ∞	2 2	2 2
$g(1) = 0$	— ∞	0 0	0 0
$h(2) = 3$	3 3	— 2	3 2
$g(2) = 3$	3 3	— 0	3 0
$h(3) = 0$	— 3	0 0	— 2
$g(3) = 2$	— 3	2 0	— 0
$h(4) = 1$	1 1	— 0	— 2
$g(4) = 1$	1 1	— 0	— 0

final sketches

解答 (2)

$$\hat{J}(d_1, d_2) = \frac{0 + 0}{2} = 0$$

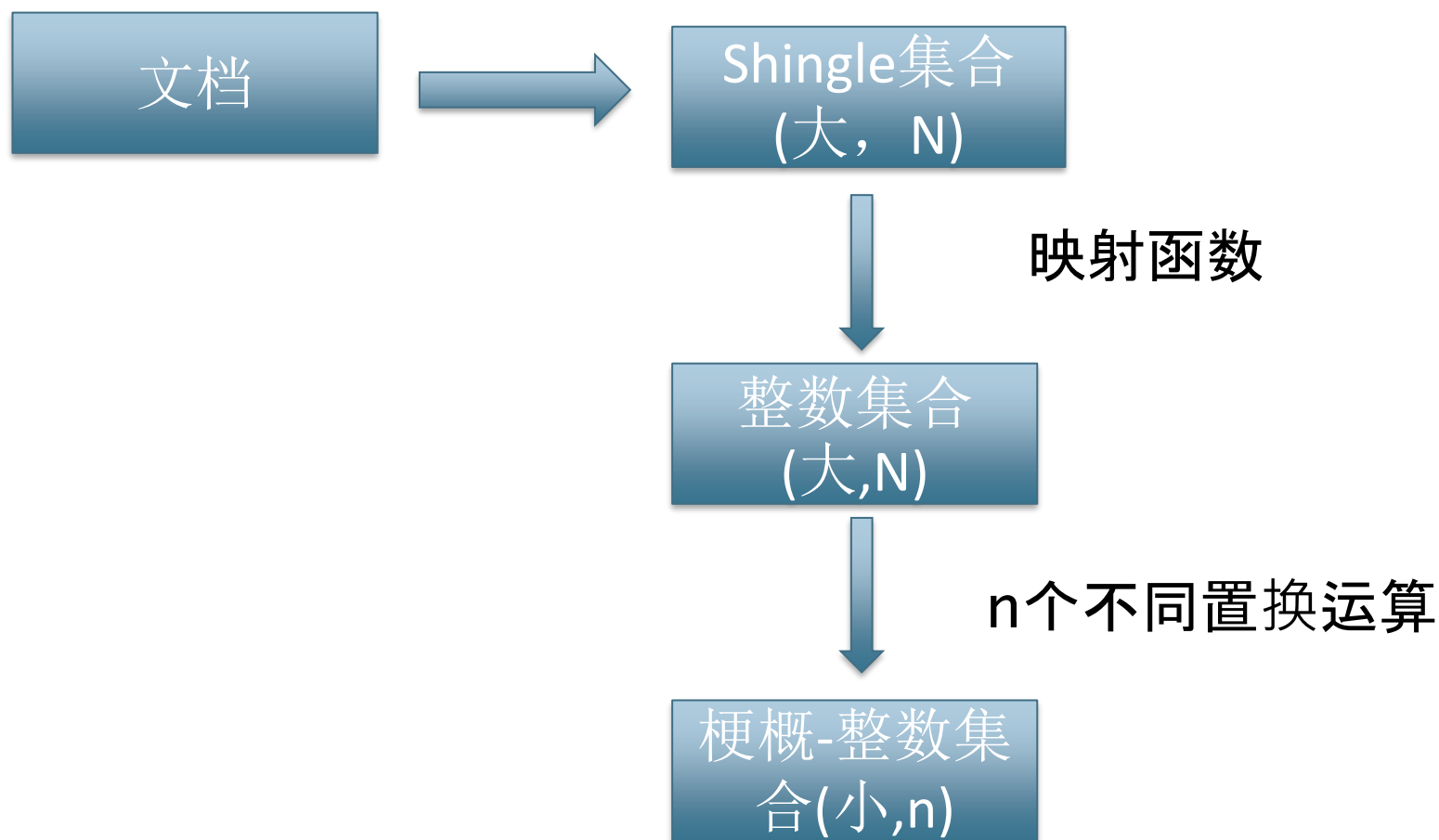
$$\hat{J}(d_1, d_3) = \frac{0 + 0}{2} = 0$$

$$\hat{J}(d_2, d_3) = \frac{0 + 1}{2} = 1/2$$

Shingling技术概要

- 输入： N 篇文档
- 选择用于生成shingle的n-gram的大小，例如 $n = 5$
- 选择200个随机置换，每个置换可以通过哈希函数表示
- 计算 N 个梗概值: 得到一个 $200 \times N$ 的矩阵(参考前面的例子)，其中每一行对应一个置换，每一列对应一个文档
- 计算 $\frac{N \cdot (N-1)}{2}$ 个两两文档之间的相似度
- 将所有两两之间相似度大于 θ 的文档构成一个传递闭包
- 对每一个传递闭包只索引一篇文档

文档的表示过程



高效的近似重复检测

- 现在我们已经得到了一个非常高效的算法来估计两篇文档的Jaccard距离
- 但是，如果Web网页数目为 N ，那么仍然需要估计 $O(N^2)$ 个相似度
- 仍然无法处理
- 一种解决办法: 局部敏感哈希(locality sensitive hashing , 简称LSH, 也常译成位置敏感哈希)
- 另一种解决办法: 排序 (Henzinger 2006)

提纲

- 互联网上的搜索
- 互联网广告
- 重复检测
- 信息爬取

采集会有多难？

- Web搜索引擎必须要采集网页文档
- 其他有些IR系统获得文档内容相对容易一些
 - 比如，对硬盘上所有文档建立索引只需要基于文件系统进行迭代式扫描即可
- 但是对于Web IR系统来说，获得文档内容需要更长的时间 ...
- ... 这是因为存在延迟
- 但是这真的是系统设计中的一个难点吗？

基本的采集过程

- 初始化采集URL种子队列;
- 重复如下过程:
 - 从队列中取出URL
 - 下载并分析网页
 - 从网页中抽取更多的URL
 - 将这些URL放到队列中
- 这里有个“Web的连通性很好”的基本假设

课堂思考题: 下列爬虫有什么问题?

```
urlqueue := (some carefully selected set of seed urls)
while urlqueue is not empty:
    myurl := urlqueue.getlastanddelete()
    mypage := myurl.fetch()
    fetchedurls.add(myurl)
    newurls := mypage.extracturls()
    for myurl in newurls:
        if myurl not in fetchedurls and not in urlqueue:
            urlqueue.add(myurl)
            addtoinvertedindex(mypage)
```

上述简单采集器的问题

- 规模问题: 必须要分布式处理
- 我们不可能索引所有网页, 必须要从中选择部分网页, 如何选择?
- 重复网页: 必须要集成重复检测功能
- 作弊网页和采集器陷阱: 必须要集成作弊网页检测功能
- 礼貌性问题: 对同一网站的访问按遵照协议规定, 并且访问的间隔必须要足够
- 新鲜度(freshness)问题: 必须要定期更新或者重采
 - 由于Web的规模巨大, 我们只能对一个小的网页子集频繁重采
 - 同样, 这也存在一个选择或者优先级问题

采集规模的数量级

- 如果要在一个月內采集20,000,000,000个页面...
- ... 那么必须要在一秒內大概采集 8000个网页!
- 由于我们采集的网页可能重复、不可下载或者是作弊网页, 实际上可能需要更快的采集速度才能达到上述指标

采集器必须做到

礼貌性

- 不要高频率采集某个网站
- 仅仅采集robots.txt所规定的可以采集的网页

鲁棒性

- 能够处理采集器陷阱、重复页面、超大页面、超大网站、动态页面等问题

Robots.txt文件

- 1994年起使用的采集器协议(即规定了采集器对网站的访问限制)
- 例子:
 - User-agent: *
Disallow: /yoursite/temp/
 - User-agent: searchengine
Disallow: /
- 采集时, 要将每个站点的 robots.txt放到高速缓存中, 这一点相当重要

Example of a robots.txt (nih.gov)

```
User-agent: PicoSearch/1.0
Disallow: /news/information/knight/
Disallow: /nidcd/
...
Disallow: /news/research_matters/secure/
Disallow: /od/ocpl/wag/
User-agent: *
Disallow: /news/information/knight/
Disallow: /nidcd/
...
Disallow: /news/research_matters/secure/
Disallow: /od/ocpl/wag/
Disallow: /ddir/
Disallow: /sdminutes/
```

<https://www.jd.com/robots.txt>

```
User-agent: *  
Disallow: /?*  
Disallow: /pop/*.html  
Disallow: /pinpai/*.html?*  
User-agent: EtaoSpider  
Disallow: /  
User-agent: HuihuiSpider  
Disallow: /  
User-agent: GwdangSpider  
Disallow: /  
User-agent: WochachaSpider  
Disallow: /
```

任意一个采集器应该做到

- 能够进行分布式处理
- 支持规模的扩展：能够通过增加机器支持更高的采集速度
- 优先采集高质量网页
- 能够持续运行：对已采集网页进行更新